IBM Cloud

Discovery v2

Product guide

Edition notices

This PDF was created on 2024-01-25 as a supplement to *Discovery v2* in the IBM Cloud docs. It might not be a complete set of information or the latest version. For the latest information, see the IBM Cloud documentation at https://cloud.ibm.com/docs/discovery-data.

Getting started with Watson Discovery

In this tutorial, we introduce IBM Watson® Discovery and walk you through the Discovery sample project. Exploring the sample project is a great way to tour and try out some of the product's features.

Before you begin

Choose the appropriate step to complete for your deployment:

- IBM Cloud Pak for Data Install Discovery. See Installing Discovery for Cloud Pak for Data .
- IBM Cloud Complete the following steps:

    1. Sign up for a IBM Cloud account or log in.
    2. You can use a Plus plan for 30 days at no cost. However, to create a Plus plan instance of the service, you must have a paid account.

       For more information about creating a paid account, see Upgrading your account.

       > ⚠ **Important:** If you decide to discontinue use of the Plus plan and don't want to pay for it, delete the service instance before the 30-day trial period ends.

    3. Go to the Discovery resource page in the IBM Cloud catalog and create a Plus plan service instance.

Step 1: Open Watson Discovery

IBM Cloud

These instructions apply to all managed deployments, including IBM Cloud Pak for Data as a Service instances.

1. Click the Discovery instance that you created to go to the service dashboard.
2. On the **Manage** page, click **Launch Watson Discovery**.

   If you're prompted to log in, provide your IBM Cloud credentials.

IBM Cloud Pak for Data

These instructions apply to Discovery deployments:

1. From the IBM Cloud Pak for Data web client main menu, expand **Services**, and then click **Instances**.
2. Find your instance, and then click it to open its summary page.

   > 🔖 **Note:** You can create a maximum of 10 instances per deployment. After you reach the maximum number, the *New instance* button is not displayed in IBM Cloud Pak for Data.

3. Click **Launch tool**.

Step 2: Open the sample project

A new browser tab or window opens and the *My Projects* page is displayed.

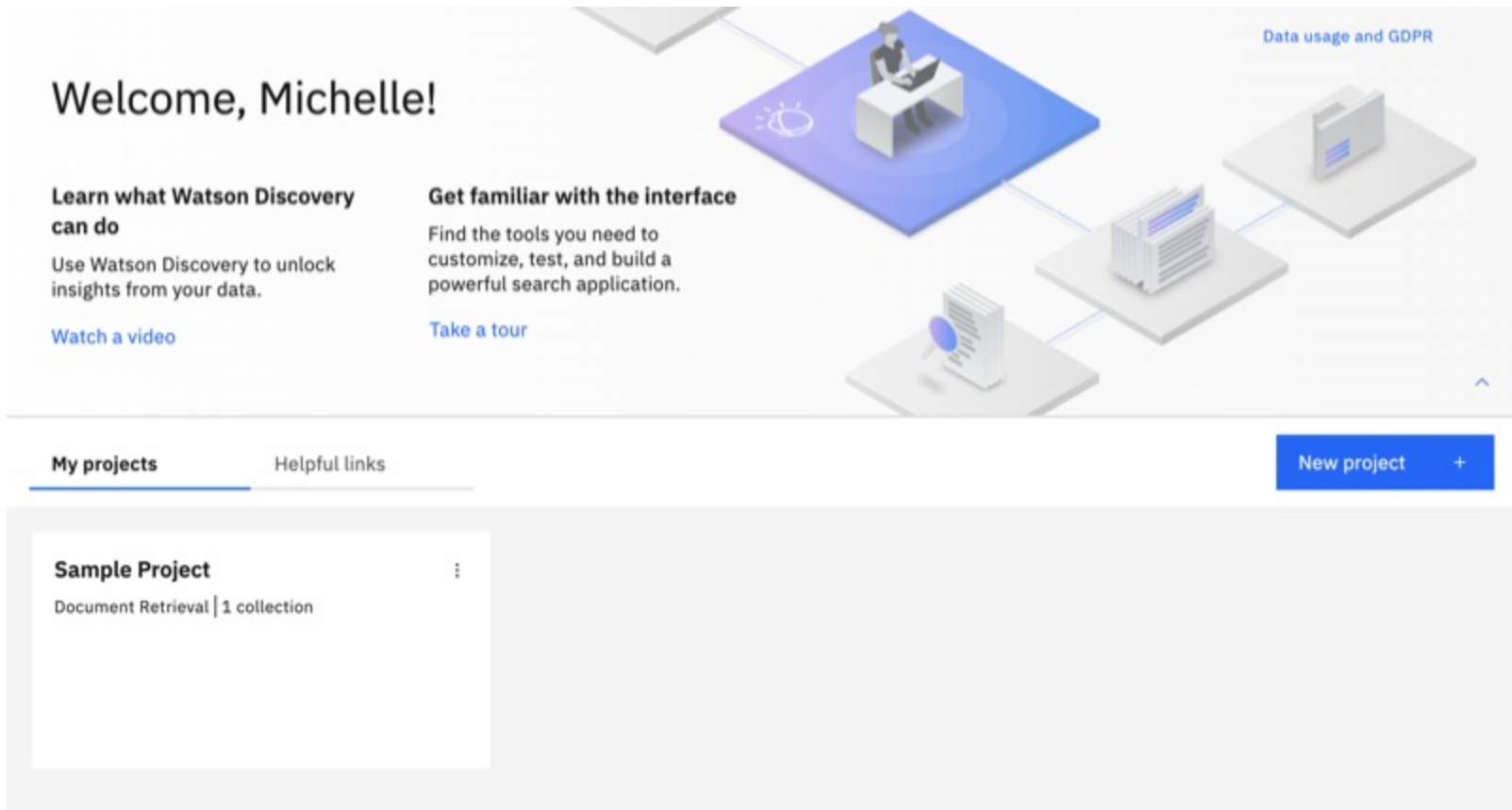

Figure 1. My projects page of the Sample project

Tip: To get familiar with the product, you can watch an under 3-minute overview video by clicking the *Watch a video* link from the product home page.

In this tutorial, you explore the sample project.

The sample project is a built-in project that is provided as a resource for you to initially explore the product. The sample project is a *Document Retrieval* project type. Document Retrieval projects are used to search and find the most relevant answers from your data.

1. Click **Sample Project**.

   The *Improve and customize* page is displayed.

   Note: If you just installed Discovery, the Sample Project needs time to finish processing documents. Wait for processing to finish before you start experimenting. You can check the status of data processing from the *Activity* page, which is described in the next step.

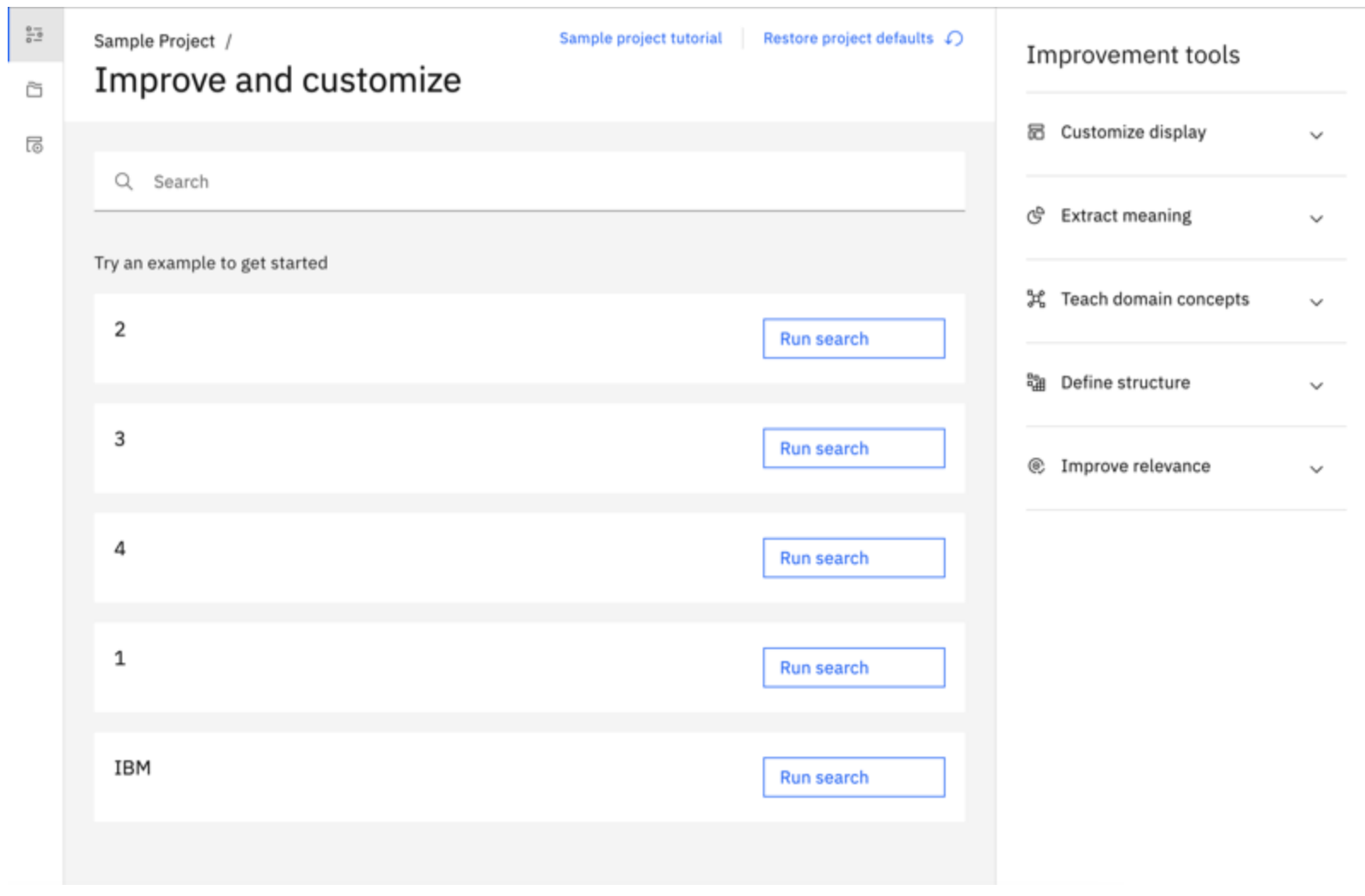

Figure 2. Sample project Improve and customize page

Step 3: Learn about the sample collection

Learn about ways you can manage and enhance a collection by exploring the sample collection that is available with the sample project. The sample collection consists of a set of uploaded IBM Support PDF documents.

1. Click the **Manage collections** icon on the navigation panel.

   Any collections in your project are displayed here. This project has only one collection.

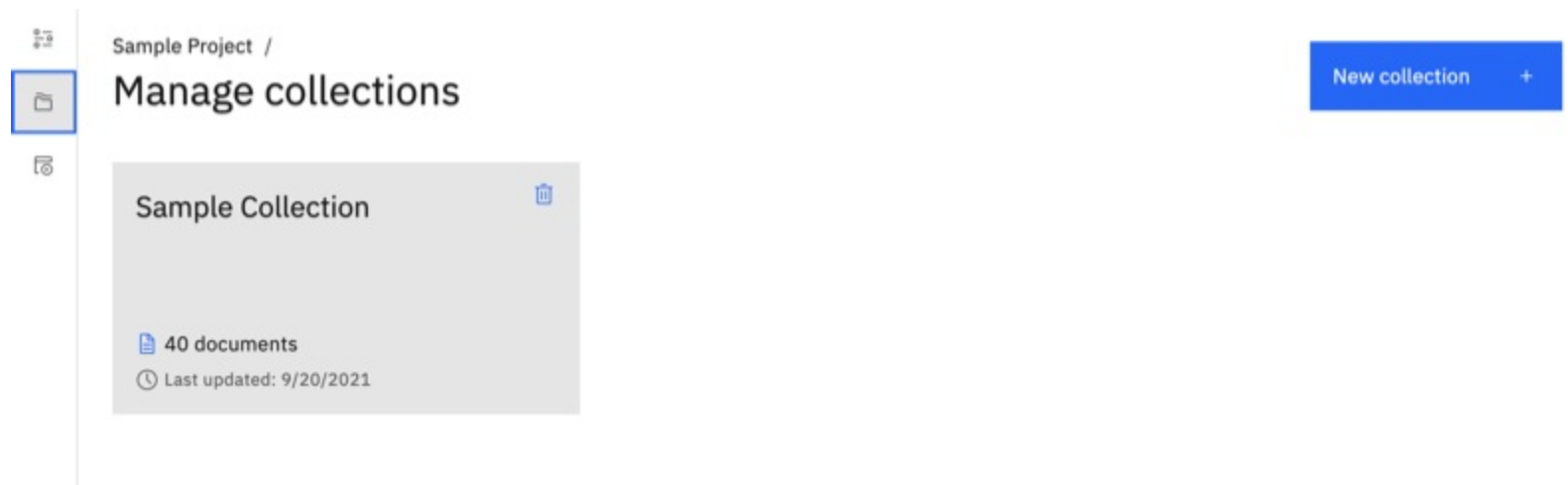

Figure 3. Collections page in the Sample project

2. Click **Sample Collection**.

   The *Activity* page is displayed. This page shows the status of the collection. For example, it shows the total number of documents and when it was last updated. If Discovery encounters a problem when a document is uploaded or a data source is crawled, any associated messages are displayed here.

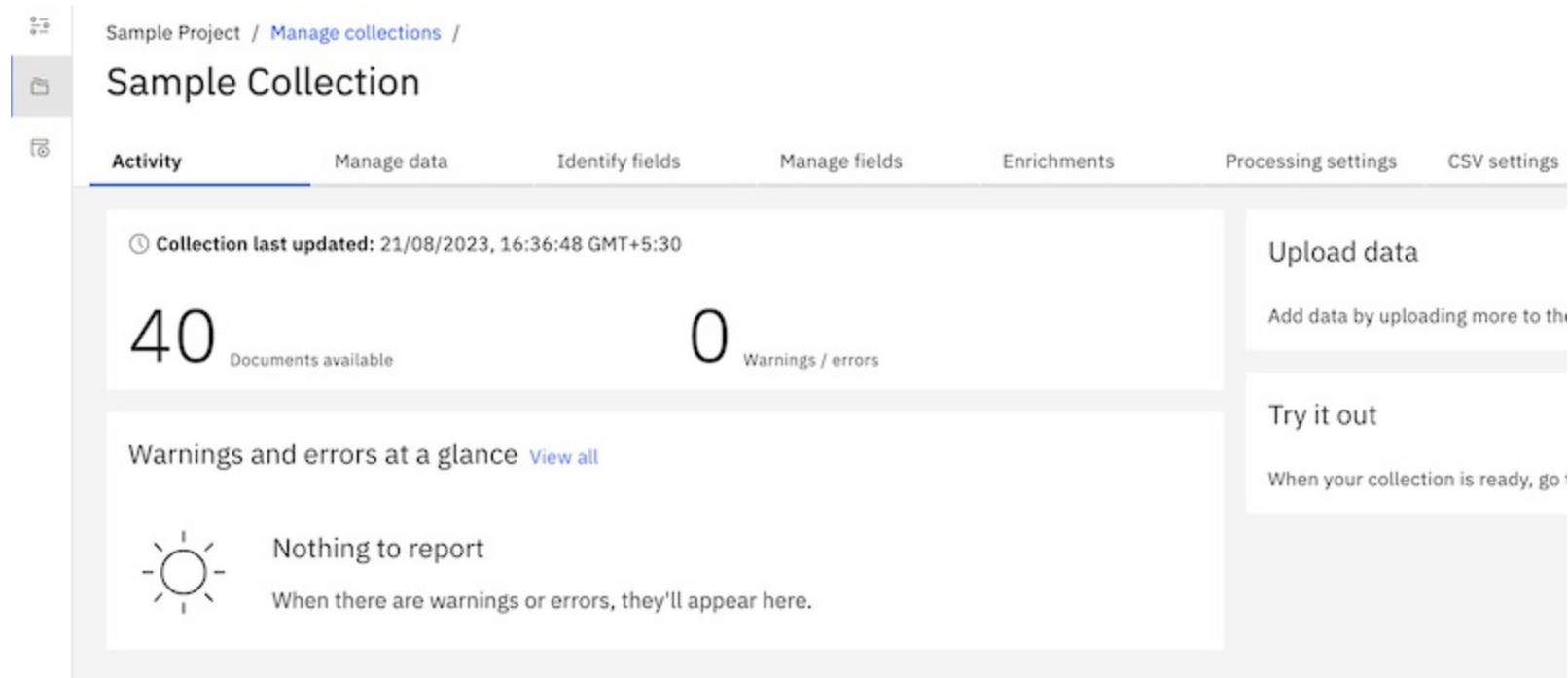

Figure 4. Activities page in the Sample project

   After you create a collection, you can come to this page to find information about the processing status of the data in the collection.

3. Click the **Enrichments** tab.

   The *Enrichments* page shows you a list of available enrichments. Enrichments make meaningful information easier to find and return in searches. You can apply built-in enrichments to your collection to leverage powerful Natural Language Understanding models that tag terms, such as commonly known keywords.

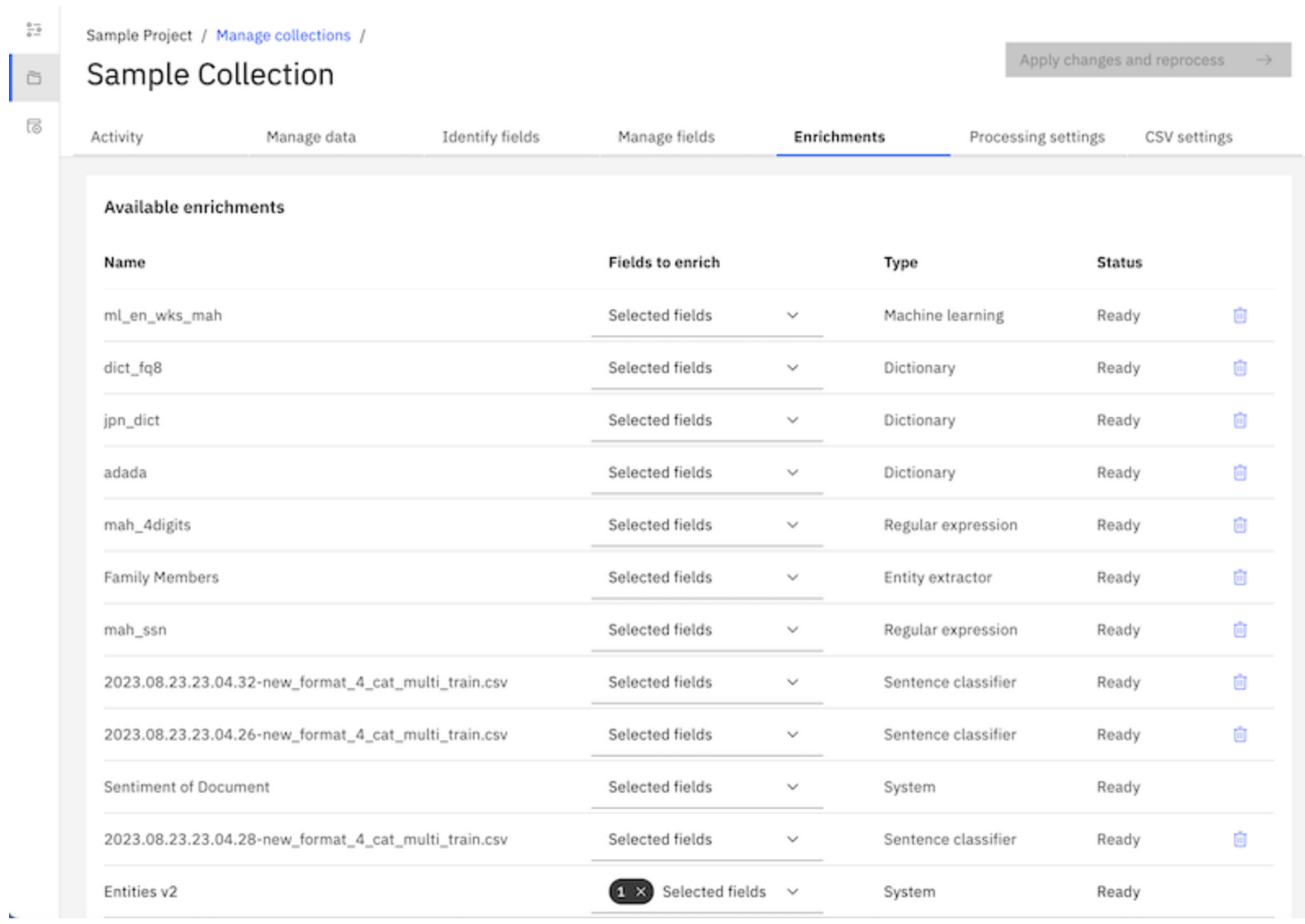

Figure 5. Enrichments page of the Sample project

The *Entities* enrichment is applied to the sample collection:

Entities

Recognizes proper nouns such as people, cities, and organizations that are mentioned in the content.

This enrichment is applied automatically to collections that are added to projects of the *Document Retrieval* type.

4. For the *Entities v2* enrichment, click **1x Selected fields**.

   A list of available fields is displayed and the `text` field is selected. This selection means that the *Entities* enrichment was applied to content that was indexed and added to a field named `text` when documents from the collection were processed.

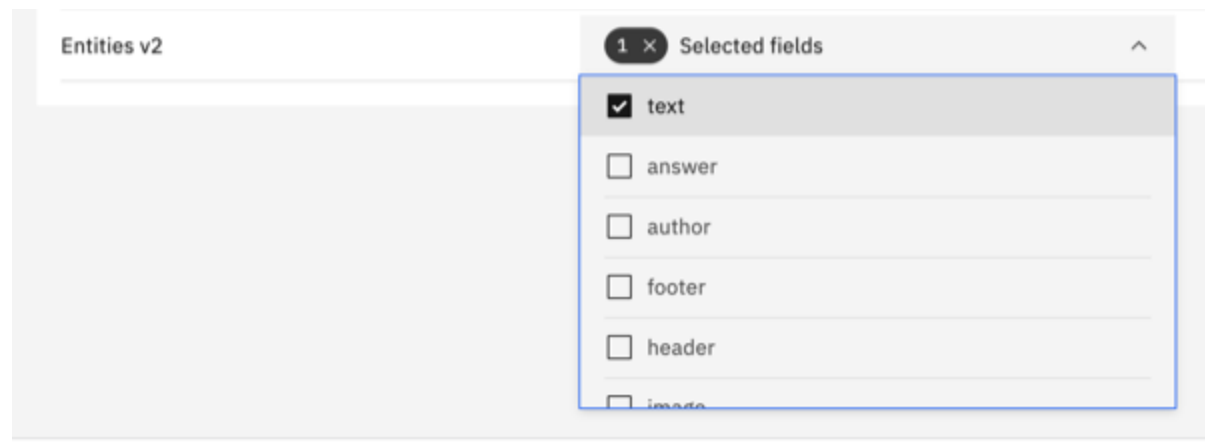

Figure 6. Entities enrichment being applied to the text field

From this page, you can apply new enrichments to your collection or change the fields where an enrichment is applied.

A powerful feature of Discovery is that you can add your own custom enrichments, such as dictionaries, patterns, and machine learning models. When you create custom enrichments, they are listed on this page also. You can manage where they are used from here.

Tip: For more information about custom enrichments, see Adding domain-specific resources.

5. You are going to apply another enrichment to the collection. Find the *Keywords* enrichment in the list, and then click **Select fields**.

   The Keywords enrichment recognizes significant commonly-known terms in your content.

6. Scroll through the list of fields until you find the `text` field, and select it.

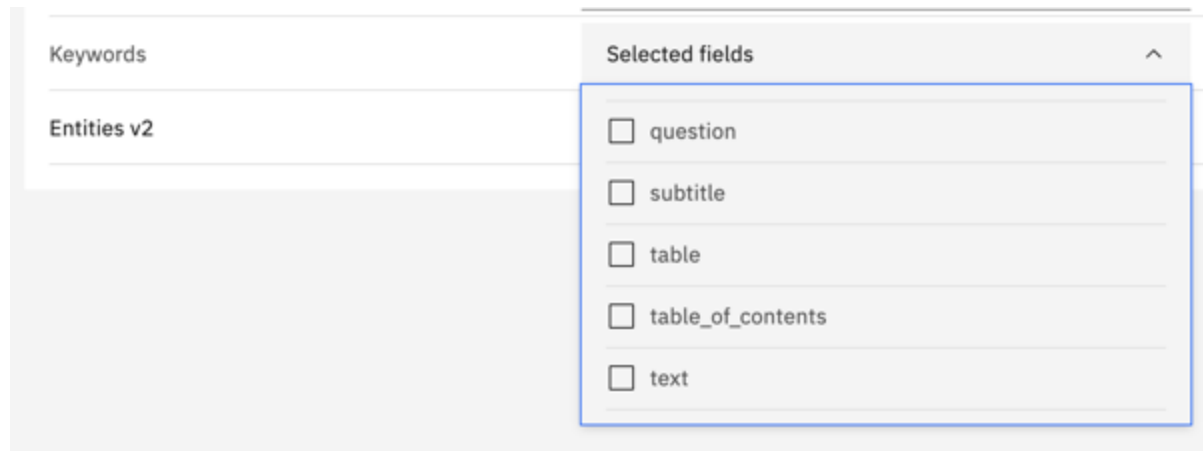

Figure 7. Fields to which you can apply the Keywords enrichment

7. Click **Apply changes and reprocess**.

   While your documents are being reprocessed to look for and tag keywords, you can continue to explore the tools available for managing a collection.

8. Click **Identify fields**.

   Most content from a document is indexed in the `text` field automatically. You might want to index certain types of content in different fields or split up large documents so that the `text` field contains fewer passages per document. To do so, you can teach Discovery to recognize important fields in your documents by applying a *Smart Document Understanding* model to your collection.

   Smart Document Understanding (SDU) is a technology that learns about the content of a document based on the document's structure. You can apply a prebuilt SDU model or create a custom SDU model.

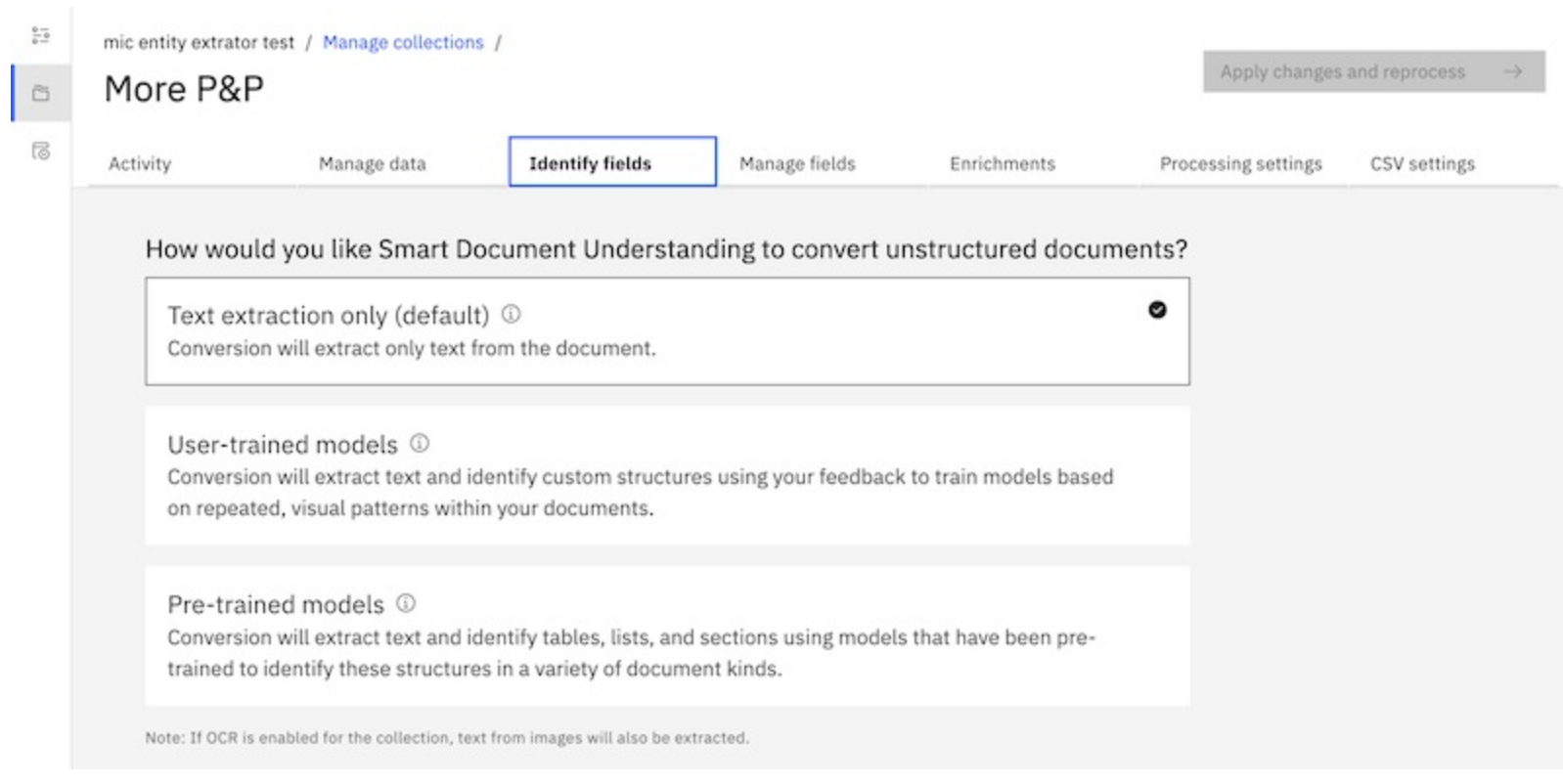

Figure 8. Smart Document Understanding model options

To create a custom SDU model, you select the *User-trained model* option, and then annotate fields in your document. (You will not annotate documents as part of this tutorial.)

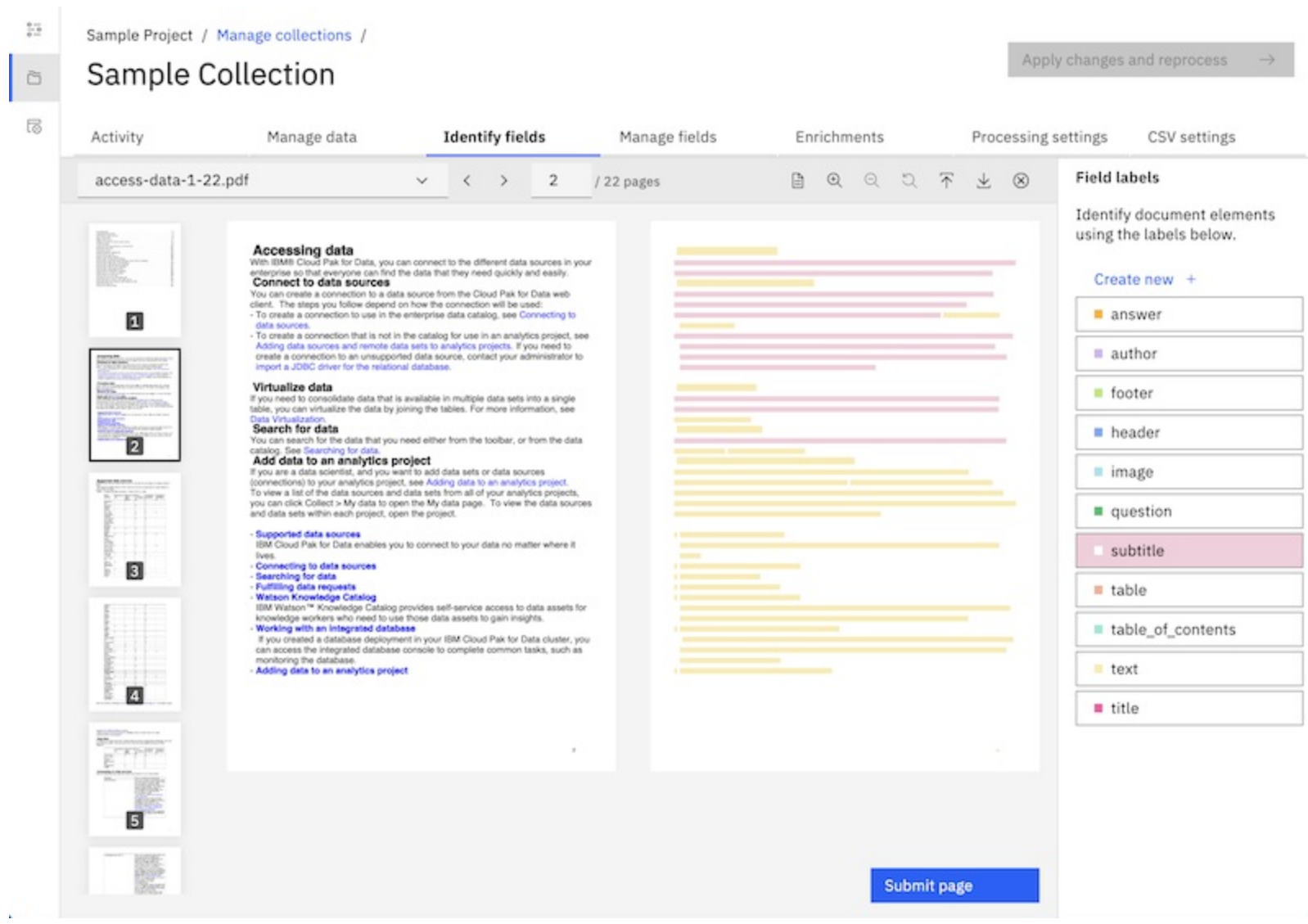

Figure 9. Smart Document Understanding annotation tool

Tip: For more information about SDU, see Using Smart Document Understanding.

9. Click **Manage fields**.

   The *Manage fields* page lists the indexed fields. From here, you can include or remove fields from the index. You can also split large documents into many smaller documents.

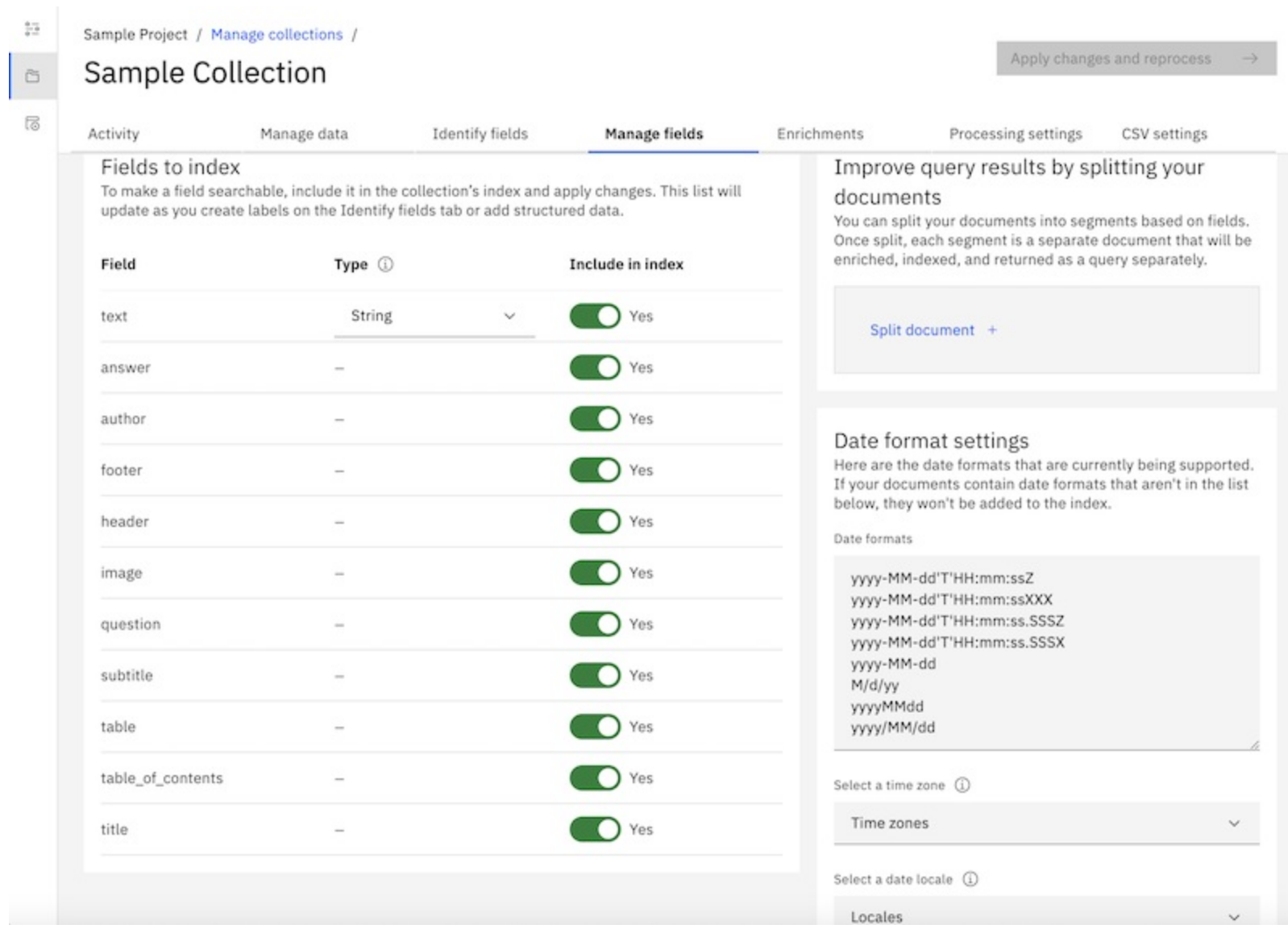

Figure 10. Fields in the collection index

Tip: For more information about splitting documents, see Splitting documents to make query results more succinct.

Step 4: Search the sample project

1. Click the **Improve and customize** icon from the navigation panel.

   The *Improve and customize* page is where you can try out queries, then add and test customizations to improve the query results for your project. A list of sample queries is displayed to help you get started with submitting test queries.

2. Click the **Run search** button for `IBM`.

   Query results are displayed.

3. From one of the query results, click **View passages in document**.

   A preview of the document where the result was found is shown.

4. Do one of the following things to explore the search result.

   1. Click **Open advanced view**.

      Useful summary information is displayed, such as the number of occurrences of any enrichments that are detected in the document.

   2. Select the `URL` entity to highlight mentions of URLs within the text.

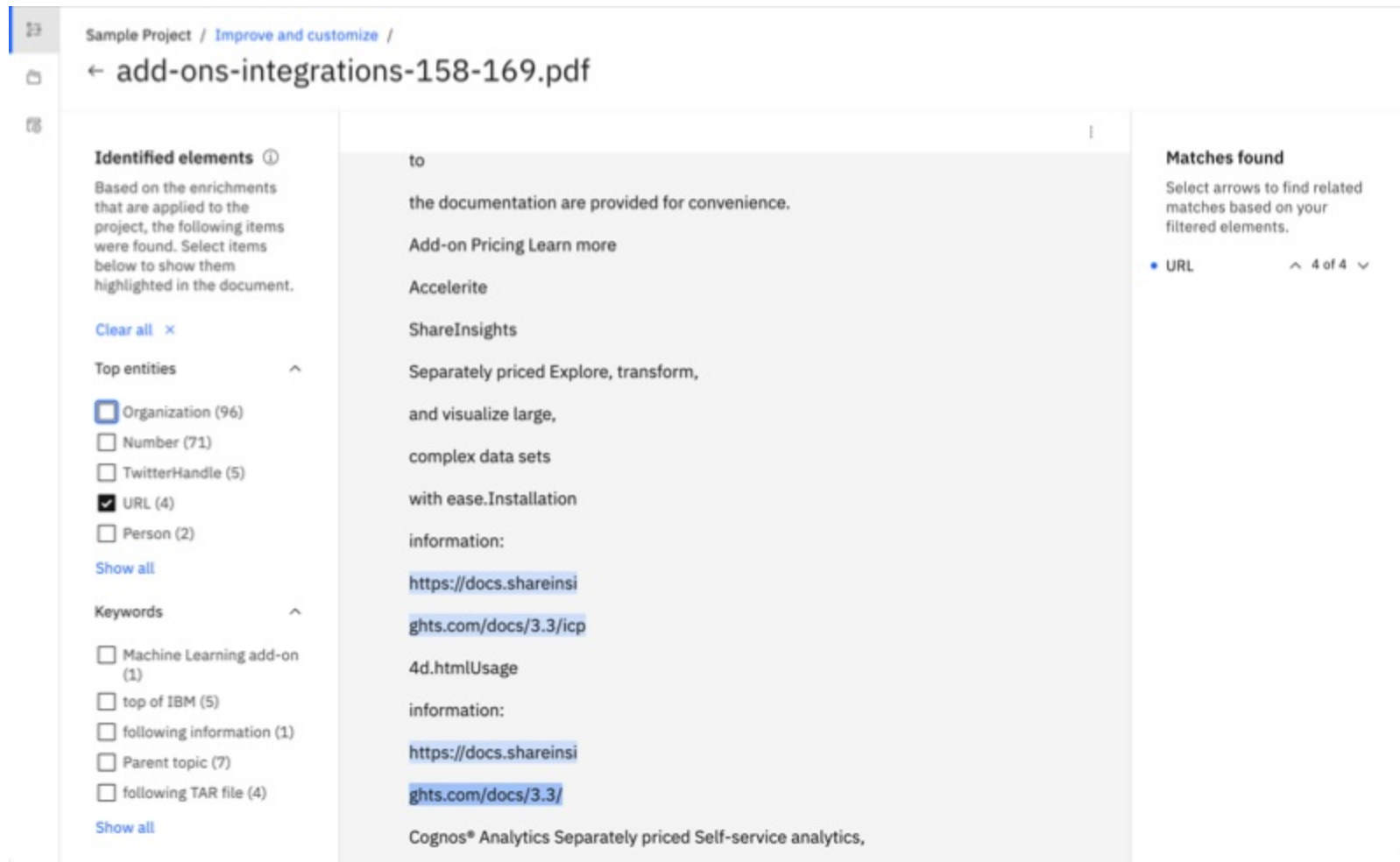

Figure 11. Advanced view that shows entities that were recognized

3. To see how the information from the document is stored in JSON format, click the *View as* menu from the view header, and select **JSON**.

   A JSON representation of the document is displayed.

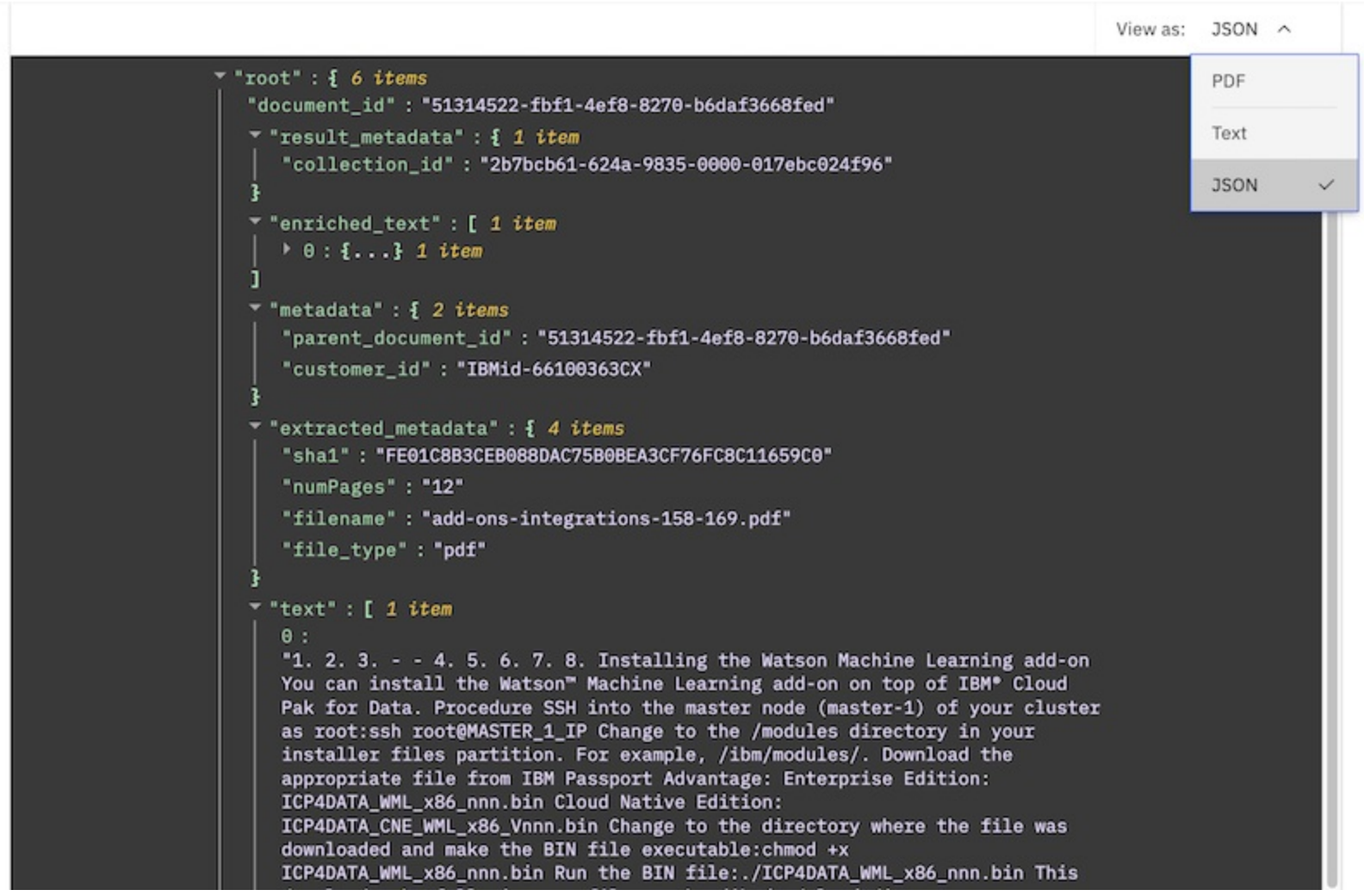

Figure 12. JSON representation of the document

You can explore the JSON representation to see information that Discovery captured from the document. For example, if you expand the `enriched_text` section, and then expand the `entities` section, you can see mentions of entities that were recognized and tagged by the Entities enrichment.

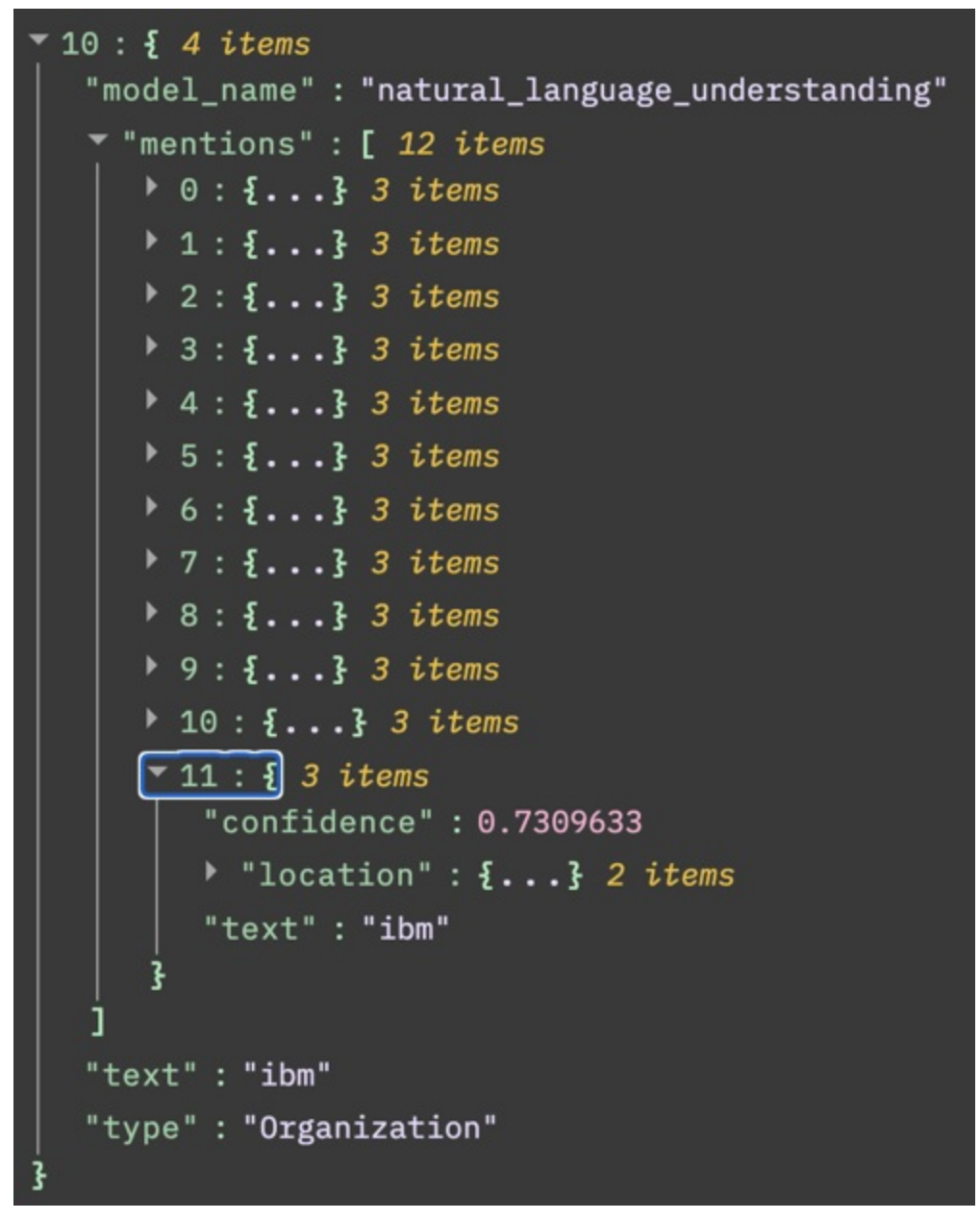

Figure 13. Shows the enrichment_text.entities section of the JSON representation

Step 5: Customize the sample project

Now, let's customize the search result view a bit by adding a facet. A facet is a way to organize and classify documents that share similar patterns or content.

1. From the *Improve and customize* page, submit the following natural language query:

```
$ How do I install Discovery?
```

2. Review the query results that are displayed.

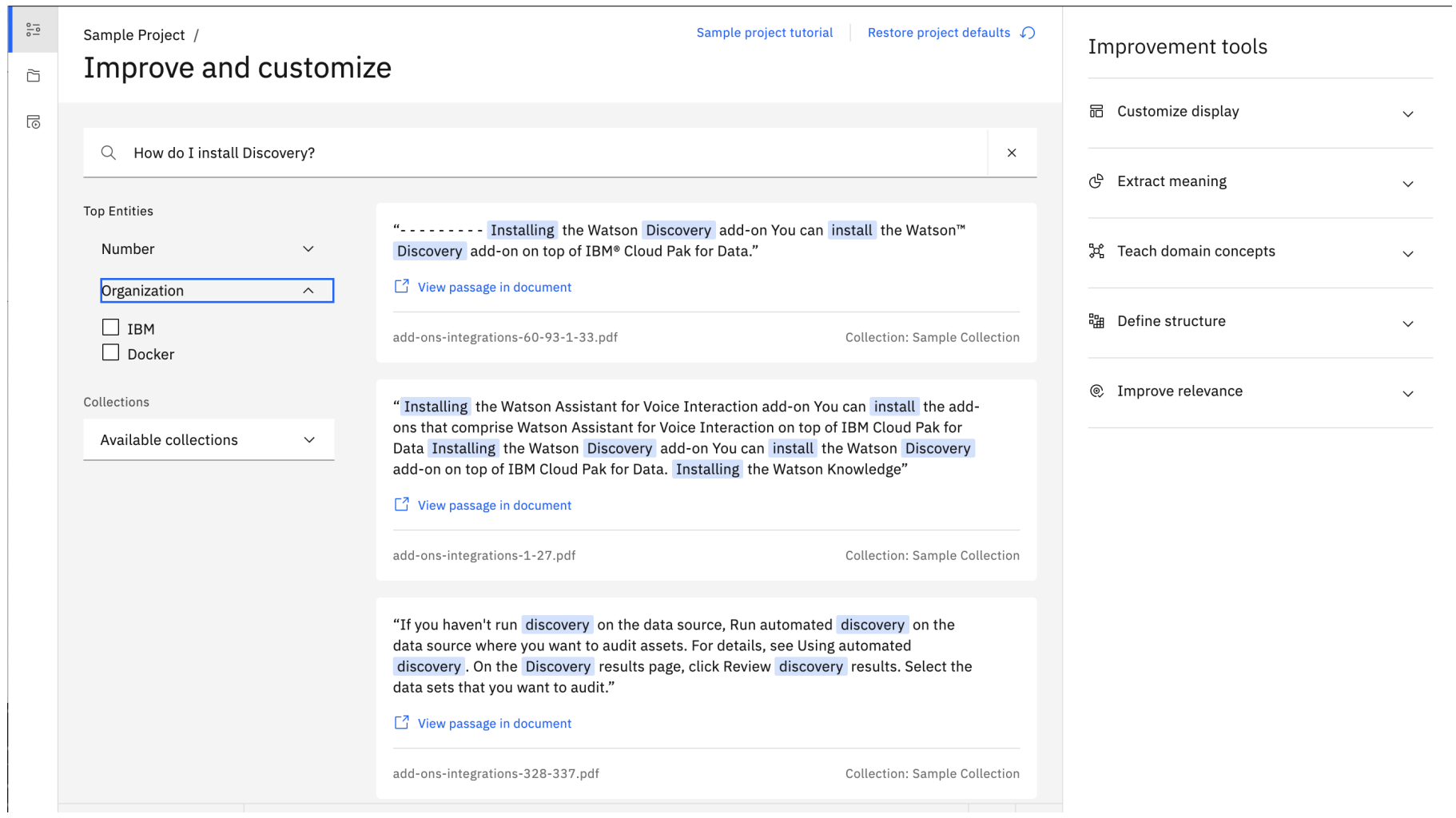

Figure 14. Top Entities facet results

Notice that a *Top Entities* section is displayed. You can expand the entities and click one of them to filter the query results to show only those results in which the entity is mentioned. The *Top Entities* section is a built-in facet. It uses information that was added to the documents by the Entities enrichment.

You will add your own facet that uses the Keywords enrichment that you applied to the collection in a previous step.

3. On the **Improvement tools** panel, expand **Customize display**, and then click **Facets**.

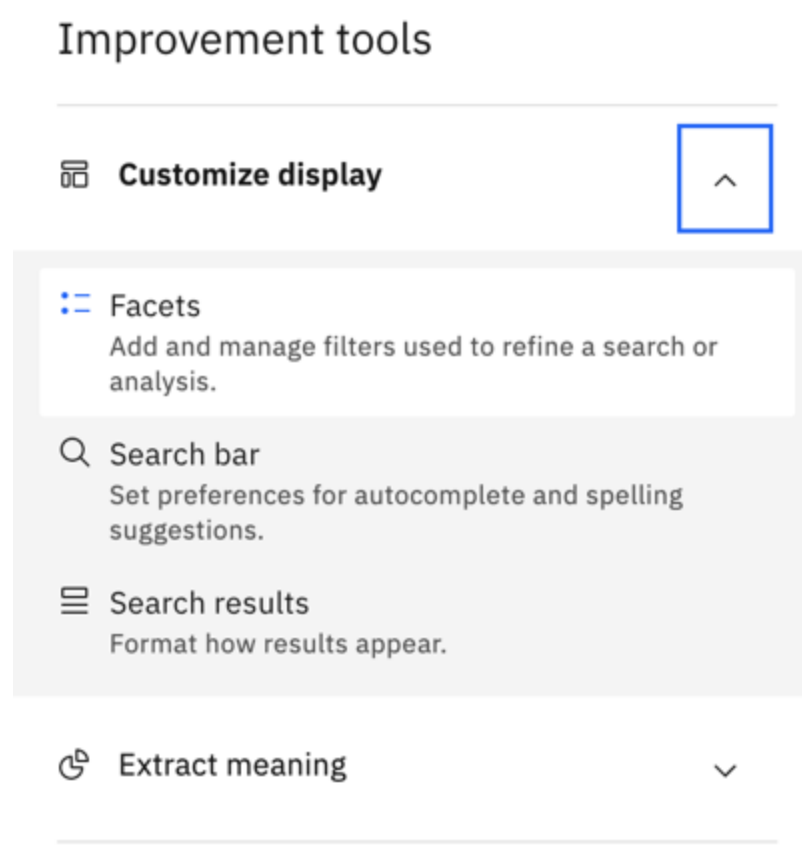

Figure 15. Customize display options

4. Click **New facet**, and then click the **From existing fields in a collection** button.
5. Choose `enriched_text.keywords.mentions.text`, change the label to `Keywords`, and then click **Apply**.

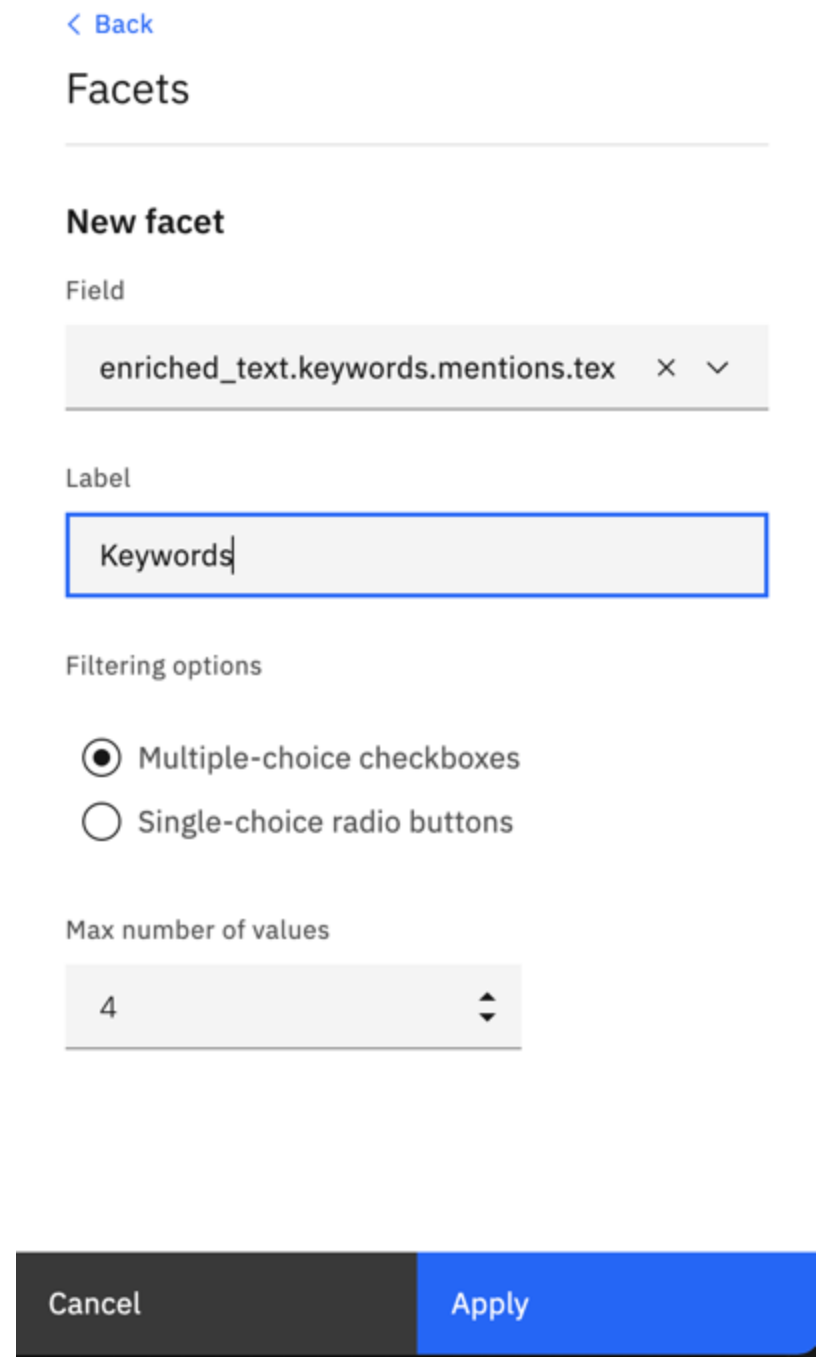

Figure 16. Creating a Keywords-based facet

Remember the JSON representation of the document that you looked at earlier? Now that the Keywords enrichment is applied to the `text` field, and the documents are reprocessed, any keyword mentions found in the `text` field are included in the JSON representation of the document.

The field you picked to use for the facet (`enriched_text.keywords.mentions.text`) reflects where the keyword text is stored in JSON.

```
"enriched_{field_name}":  [
   "keywords"  :  [
      "mentions"  :  [
         "text":  "Cloud Pak"
      ]
   ]
]
```

6. The new facet is displayed. You can click a keyword to filter the documents to include only those results that mention the keyword.

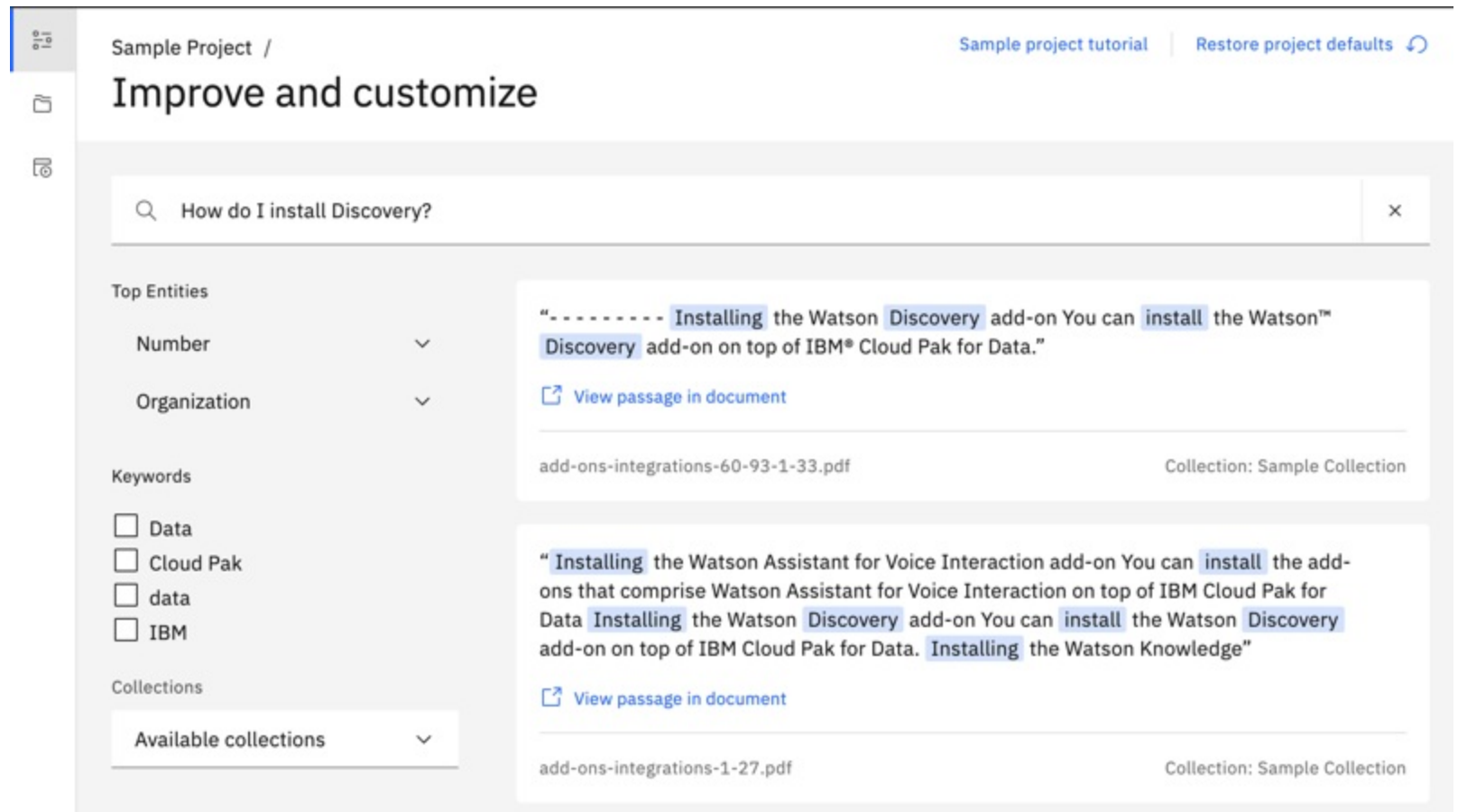

Figure 17. Keywords facet

You successfully added a built-in NLU enrichment that recognizes keywords in the sample collection documents. Then, you added a facet that uses the keywords enrichment to let you filter the documents by keyword.

Step 6: Share the sample project

1. Click **Integrate and deploy** from the navigation panel.

   From here, you can share your project with colleagues and deploy it.

2. Follow the on-screen instructions to add a user, and then send login credentials and the provided link to your colleague.

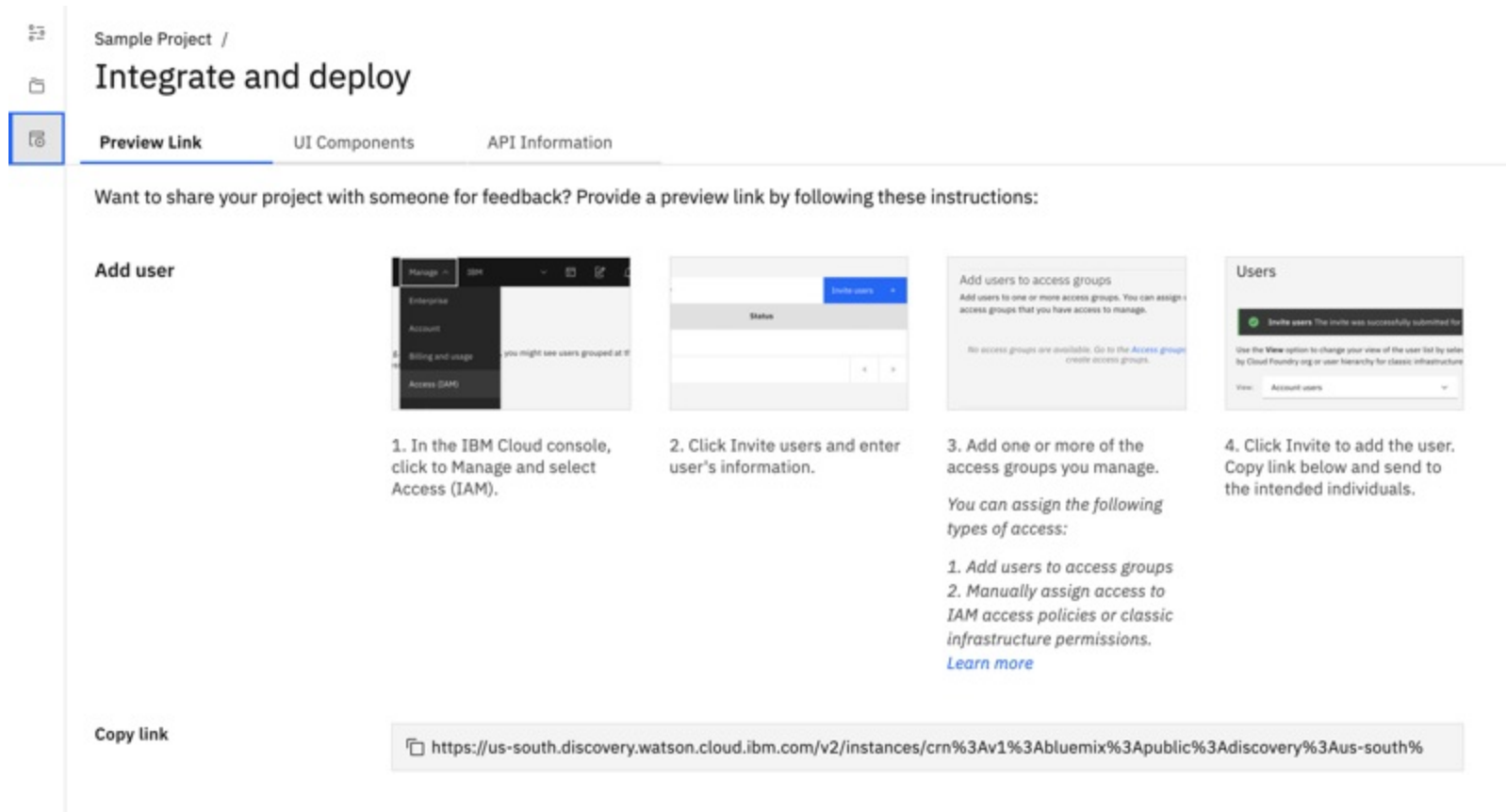

Figure 18. Integrate and deploy page

After you build your own search application and are ready to deploy it, you can use prebuilt user interface components or build a custom application.

- Click **API Information**. From this page, you can get the project ID for your project. You need the project ID to use the Discovery API. You also need the service instance URL and API key. The credential details are available from the Manage page of your service instance in IBM Cloud.
- Click **UI Components** to find links to ready-to-use code that you can use to create a full-featured search application faster.

Step 7: Add your own content

Now that you know more about some of the product features, you're ready to evaluate the data you want to search.

It's all about the data. Review the types of content you own that you want your search solution to be able to leverage.

Supported data sources

The following table shows the supported data sources for each deployment type.

Data source	IBM Cloud	IBM Cloud Pak for Data
Box	✓	✓
Database (IBM Data Virtualization, IBM Db2, Microsoft SQL, Oracle, Postgres)		✓
FileNet P8		✓
HCL Notes		✓
IBM Cloud Object Storage	✓	
Local file system		✓

Salesforce	✓	✓
Microsoft SharePoint Online	✓	✓
Microsoft SharePoint On Premises	✓	✓
Website	✓	✓
Microsoft Windows file system		✓

Supported data sources

Step 8: Not sure what you can build?

For more information about the types of search solutions you can build, see Start getting value from your data.

Tip: You can access the product documentation at any time by selecting the Help icon from the page header of the product user interface. The help content is customized to provide information that is related to what you're doing in the product.

No matter what you build, step one is to create a project. Decide which project type best fits your needs.

If none of the existing types is quite right, you can choose **None of the above** to create a custom project instead.

Project descriptions

Need	Goal	Project type
Which document contains the answer to my question?	Find meaningful information in sources that contain a mix of structured and unstructured data, and surface it in a stand-alone enterprise search application or in the search field of a business application.	**Document Retrieval**
Where is the part of the contract that I need for my task?	Quickly extract critical information from contracts.	**Document Retrieval for Contracts**
I want the chatbot I'm building to use knowledge that I own.	Give a virtual assistant quick access to technical information that is stored in various external data sources and document formats to answer customer questions.	**Conversational Search**
I want to uncover insights I didn't know to ask about.	Gain insights from pattern analysis or perform root cause analysis.	**Content Mining**

Project type use cases

For more information, see Creating projects.

About Watson Discovery

IBM Watson® Discovery is an intelligent document processing engine that helps you to gain insights from complex business documents.

Use Discovery to visually train AI for deep understanding of your content, including tables and images, to help you find business value that is hidden in your enterprise data. Use natural language or structured queries to find relevant answers, surface insights, and build AI-enhanced business processes anywhere.

Start by connecting your data to Discovery. Next, teach Discovery to understand the language and concepts that are unique to your business and industry. Enrich your data with award-winning Watson Natural Language Processing (NLP) technologies so you can identify key information and patterns. Finally, build search solutions that find answers to queries, explore your data to uncover patterns and insights, and leverage search results in automated workflows.

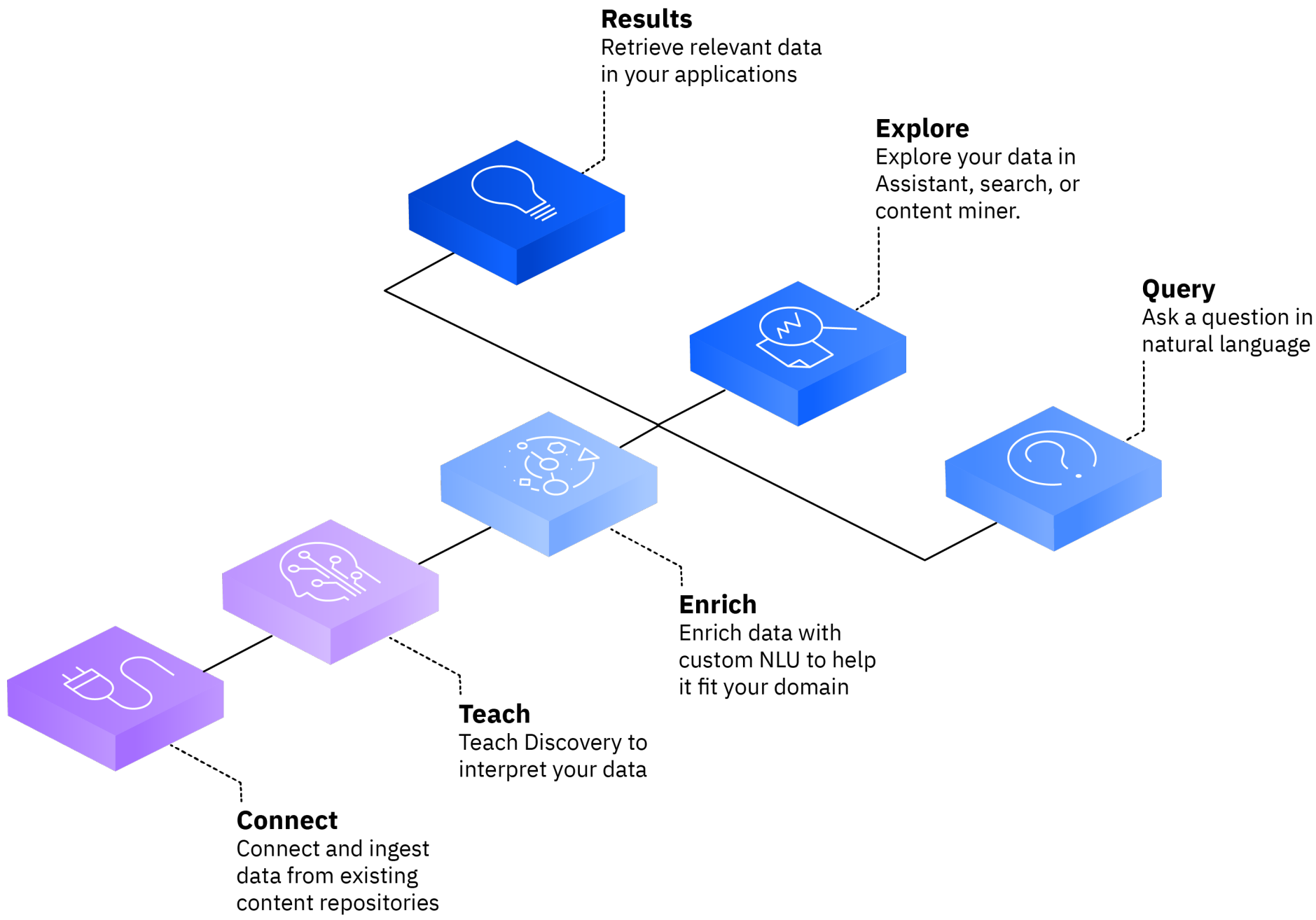

Figure 1. How to use Watson Discovery

Find out how Discovery is transforming data into artist insights at the 2023 GRAMMYs®. Read the IBM Business Operations blog post to learn more.

Overview video

Watch a video about how Discovery uses AI-powered search, retrieval, and content mining. This overview covers the key basics of projects, collections, fields, and enrichments. It explains how to upload your data and query for answers, find insights, and spot trends.

View video: Get started with Watson Discovery

Video transcript

Get started with Watson Discovery presented by David Williams - (Music intro) Welcome to Watson Discovery with AI.

In this video, we'll walk through some key concepts and show you how to get started.

Watson Discovery is made up of four main concepts, projects, collections, fields, and enrichments.

A project is a space where can import different types of data from a variety of sources, and query for insights or answers.

A collection is a set of documents that you upload or crawl from a connected data source.

As documents are crawled, unstructured text is organized into fields such as author, file type, text, and more.

And enrichments are AI capabilities that you can apply to fields to identify and extract relevant information from your documents. This helps you find answers or insights from your data.

Let's dive in to the different project types.

A document retrieval project is used to build an AI-powered search function that finds answers in your business data.

A conversational project is used to enhance your chatbot's question and answer ability.

A content mining project helps you spot trends across large volumes of text-heavy business data.

Watson Discovery supports a wide selection of data sources you can crawl, like webpages, Cloud Object Storage, Microsoft SharePoint, and more. You can even upload your own data from any data source.

After connecting and processing your data, you can apply enrichments to bring your data to life. Some commonly used enrichments are entities, contracts, and table understanding. Entities enrichment can be used to recognize people, organizations, and more. Contracts enrichment can be used to decompose contracts to fields, clauses, and relationships. The table understanding enrichment can be used to identify tables and return them as an answer to a query.

You can also create custom enrichments, such as a dictionary, so Discovery can understand your industry-specific terminology and support intelligent queries.

Now, you know the basics.

To get started, take our step-by-step product tour to get familiar with the user interface and sample project.

Using Discovery

Discovery can be deployed as a managed cloud service or can be installed on premises. This documentation describes how to use the product regardless of how it is deployed. Information that applies exclusively to one deployment type is denoted by the appropriate icon:

- IBM Cloud Pak for Data for installed instances, such as IBM Watson® Discovery Cartridge for IBM Cloud®.
- IBM Cloud for managed instances, such as Discovery Plus, Enterprise, and Premium plan instances that are hosted by IBM Cloud or instances that are provisioned with IBM Cloud Pak for Data as a Service .

Tip: Click the Help icon from the header of any page in the product user interface to open the Discovery documentation.

Browser support

IBM Cloud Pak for Data

- The minimum required browser software for the product user interface includes the following browsers:

Google Chrome

Latest version -1 for your operating system

Mozilla Firefox

Latest regular -1 and Extended Support Release (ESR) version for your operating system

Microsoft Edge

Latest version -1 for Windows

Apple Safari

Latest version -1 for Mac

- The IBM Cloud Pak for Data web client where you create service instances supports the IBM Cloud Pak for Data requirements. For more information, see Supported web browsers

IBM Cloud

- Deployments of Discovery that are managed by IBM Cloud follow the IBM Cloud requirements. For more information, see Prerequisites
- For more information about browser support for deployments that are provisioned with Cloud Pak for Data as a Service, see Which web browsers are supported for Cloud Pak for Data as a Service.

Language support

Language support by feature is detailed in the Supported languages topic.

Beta features

IBM releases services, features, and language support for your evaluation that are classified as beta. These features might be unstable, might change frequently, and might be discontinued with short notice. Beta features also might not provide the same level of performance or compatibility that generally available features provide and are not intended for use in a production environment.

Terms and notices

IBM Cloud

- IBM Cloud Terms of use
- Service terms (Search for Watson Discovery)
- Data Processing and Protection Datasheet

IBM Cloud Pak for Data

- Security on Cloud Pak for Data

Trademarks are listed in the Trademarks page for all IBM Cloud services.

Getting the most from Discovery

Getting the most from Discovery

Discovery was redesigned to introduce new features and a simpler way to build solutions.

The redesigned product is referred to as Discovery v2. When you create an instance on IBM Cloud or install and provision an instance on IBM Cloud Pak for Data, you get the new and improved version of Discovery.

Advantages of using the latest version

Discovery v2 offers the following features and enhancements:

- A project-based experience that supports many different use cases within a single environment.
- Built-in customization tools for adding dictionaries, patterns, and classifiers to help business users build projects that understand the language of their domain.
- Connectors to popular data sources that can quickly access valuable data where it resides.
- Smart Document Understanding that learns from the structure of human-readable documents, such as PDFs.
- Natural language query support across all document types, optimized with machine learning to find targeted answers.
- Advanced search capabilities, such as answer finding, curations, and table retrieval.
- An out-of-the-box contract understanding function that helps you search and interpret legal contracts.
- A full-featured Content Mining application that you can use to conduct in-depth analysis of unstructured text.
- Customizable user interface components that help you to deploy custom applications.

For more information, see Migrating to Discovery v2.

Comparing v1 and v2 features

If you are already familiar with Discovery v1, learn more about how Discovery v2 compares.

Discovery v2 has new features that were previously unavailable. The following table describes feature support in both versions.

Feature	Product redesign (v2)	Earlier version (v1)
Use projects to organize your work	✓	
Use the Smart Document Understanding (SDU) to annotate your documents	✓	✓
Leverage intuitive user interface tools to add domain-specific artifacts, such as dictionaries and custom machine learning models	✓	
Create a content mining project type and then use the built-in Content Mining application to do in-depth data analysis (*IBM Cloud Pak for Data, Enterprise, and Premium plans only* **)**	✓	
Perform real-time NLP with the Analyze API (*IBM Cloud Pak for Data and Enterprise plans only* **)**	✓	
Apply a pretrained Smart Document Understanding model to your collection for similar benefits with less effort	✓	
Process text from scanned documents or other images	✓	✓
Extract meaning from tables	✓	
Get insights from contracts (*IBM Cloud Pak for Data, Enterprise, and Premium plans only* **)**	✓	
Apply the *Part of Speech* **enrichment to your data**	✓	
Use the Entity Extraction, Document and Phrase Sentiment Analysis, and Keyword Extraction enrichments	✓	✓
Use the Category classification, Concept tagging, Relation Extraction, Emotion Analysis, and Semantic Role Extraction, Sentiment of Keywords and Entities enrichments, which are available with the Natural Language Understanding service		✓

Build a custom entity type system	✓	
Apply Watson Knowledge Studio NLP models to your data	✓	✓
Support for more connectors from a IBM Cloud Pak for Data deployment, including databases, file systems, FileNet P8, and HCL Notes	✓	
Some connectors support document-level security from a IBM Cloud Pak for Data deployment	✓	
Programmatically configure external data source crawls		✓
Configure the normalization processes of document segmentation and HTML file inclusion or exclusion rules during ingestion		✓
Configure the JSON normalization process during ingestion and after enrichment	✓	✓
Configure dictionary tokenization		✓
Advanced question-answering capabilities, such as returning the exact answer	✓	
Discovery Query Language (DQL) API support	✓	✓
Retrieve passages from documents	✓	✓
Perform relevancy training to improve query results	✓	✓
Configure continuous relevancy training		✓
Retrieve tables	✓	
Query result deduplication		✓
Identify document similarity in query results	✓	✓
Indicate a preference (bias) in queries		✓
Review query logging and metrics		✓

Feature support details

Limit details

For more information about artifact limits per plan, see the feature documentation:

- Advanced rules model limits
- Classifier limits
- Collection limits
- Dictionary limits
- Document limits
- Entity extractor limits
- Machine Learning model limits
- Pattern limits
- Project limits
- Query limits
- Regular expression limits
- SDU limits

The following limits apply only to Content Mining project types:

- Document classifier limits
- Regular expression pattern limits

To check the current status of the limits and usage for your plan type, you can open the *Plan limits and usage* page at any time.

1. From the product page header, click the user icon .

   The *Usage* section shows a short summary.

2. Click **View all** to see usage information for all of the plan limit categories.

To leave the page, click the web browser back button or the *My Projects* tab.

Migrating to Discovery v2

A redesign of the product, Discovery v2, was introduced in November 2019. Discovery v2 offers significant advantages over Discovery v1.

Learn about how to migrate a v1 Discovery service instance to Discovery v2, including how to move data and update your applications.

The major structural differences between Discovery v1 and v2 include:

- There is no concept of an environment in v2. The deployment details such as size and index capacity are managed for you when you choose the appropriate service plan for your needs. For managed deployments, you can choose a Plus, Enterprise, or Premium plan, for example. For installed deployments, the sizing is managed by the deployment type that you specify when you install the service in Cloud Pak for Data.
- There is no single configuration object in v2. Control of the enrichments that are applied to documents is managed in the collections and project objects in v2. Other v1 configuration capabilities, such as the ability to customize the conversion step of ingestion, are not available in v2.
- Greater programmatic support is available for custom enrichments in v2. New enrichment API methods are available that you can use to create enrichments. v2 also introduces document classifier API methods that you can use to train document classifier models programmatically. You can apply these custom enrichments to a collection by using the API.
- The capabilities of a natural language query search are expanded in v2 to enable the return of the top passages per document and of succinct answers from passages. Other advanced search capabilities are introduced, including table retrieval. In v2, the deduplication parameter is not available and the continuous relevancy training and query logging functions are not available.
- For more information about feature differences, see the feature comparison table.
- For more information about detailed API differences, see API version comparison.

Discovery v2 is available for all users of Plus or Enterprise plan instances, or Premium plan instances that were created after 15 July 2020. v2 is also available for IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data users.

Migration overview

Migrating from Discovery v1 to v2 is a multistep process that you can do independently.

The two versions of the Discovery service have many differences, but you can adopt techniques and utilities that were applied to a v1 instance for use with your new v2 instance.

To migrate from v1 to v2, you must complete the following high-level steps:

1. Plan the migration.
2. Transfer your documents.
3. Update your application to use the v2 API.
4. Regression test and deploy the updated application.
5. Delete your v1 plan service instance.

Note: Some steps require you to make programmatic changes by using the API and others involve changes that you can make from the product user interface.

Plan the migration

Get familiar with what's new in v2 and learn about how it differs from v1 before you provision a v2 instance. Your first v2 Plus plan trial instance is available at no charge for 30 days. Learn about and plan for the migration before you provision the instance so that you can get the most from your trial.

When you're ready to start the migration, create a migration schedule that you and your team can follow as you complete the process. Be sure to set up the new v2 service instance and get projects and collections re-created in the new service instance before you switch over to using the v2 service and before you delete your v1 instance.

Learn about the Discovery v2 plan options, so you can choose the right plan for your long-term needs. The Plus plan that you use to get started might be sufficient. However, you might choose to use an Enterprise or Premium plan instead. From a Plus plan, you can do an in-place upgrade to an Enterprise plan, but not to a Premium plan.

Plan how to adapt your application

One of the main changes between versions is that Discovery v2 introduces projects. A project consists of one or more collections. The advantage of using projects is that one query can run against many collections at the same time. Each collection can contain documents that you upload or that you crawl from a single data source, such as a website, Microsoft SharePoint, and more.

Things to consider when you adapt your application to use projects:

- Although the concept of an environment does not exist in v2, data is still organized into collections. In v2, collections are grouped into projects. In most cases, you want to migrate a single v1 collection to a single v2 collection.

  If you want to keep relevancy training information that is applied to a v1 collection, add the collection documents to a single collection in your v2 project.

- Decide how many collections you want to add to each v2 project. All project types, except Content Mining projects, can contain up to 5 collections. Choose the right type of project for your data.

  To optimize search results, different enrichments and configuration options are applied automatically to collections that are added to different project types. For more information, see the following topics:

  - Project descriptions
  - Default project settings
  - Default query settings

- The Discovery v2 API changed to account for projects and collections, among other enhancements. Some API calls changed to support actions at the project level instead of the collection level, such as submitting a query and running relevancy training. Many other API methods changed and some are not available in v2. For a detailed comparison of the v1 and v2 API methods, see API version comparison.

Picking a service plan

Choose among the *Plus*, *Enterprise*, and *Premium* managed plans or opt for an on-premises installation by purchasing the Discovery Cartridge for IBM Cloud Pak for Data. Review the benefits and limits of each type of plan before you choose one.

- For more information about the plans, see Discovery pricing plans.
- For more information about artifact limits, see Limit details.

The following table shows plan types for managed deployments that are generally similar between v1 and v2.

Current v1 plan	Example v1 data usage	Similar v2 plan
Lite	Not applicable	Plus Trial (no charge for 30 days only)
Advanced (low usage)	10,000 documents, 10,000 queries per month	Plus
Advanced (high usage)	100,000 documents, 100,000 queries per month	Enterprise
Premium	Not applicable	Enterprise or Premium

Similar plans

Tip: To get information about the current storage, documents, and collections used, click the *Environment details* icon from the product user interface header.

You cannot do an in-place upgrade from a v1 plan, such as Lite or Advanced, to a v2 plan. You must create a new v2 plan, and then move your data to the new service instance. While you migrate your data from v1 to v2, you will likely have both a v1 and v2 instance deployed at the same time. Consider using the 30-day no charge trial that is available with your first Plus plan instance during this time.

Collecting metrics

Make a note of the following information so you can compare it to your service instance data after the migration:

- Number of collections

  To get the number of collections in an instance in v1, use the List collections API.

- Number of documents per collection

  To get the number of documents in a collection in v1, use the Get collection details API.

  ```
  GET {url}/v1/environments/{environment_id}/collections/{collection_id}`
  ```

The API returns information about the status of the documents in the collection, which includes the total number of available documents.

```
"document_counts":  {
       "available":  34,
       "{other}":"{values...}"
}
```

Transferring documents from v1 to v2

How you transfer your documents depends on the technique that was used to ingest the documents in v1.

Re-create one collection at a time. If you start multiple ingestion processes at the same time, you can tax the system resources and increase the overall time that it takes for the processing to be completed. You also want to keep an eye out for any informational messages that are generated by the ingestion process. It is easier to troubleshoot an ingestion issue, for example, when you ingest one collection at a time.

Uploaded data

If you used the API to upload documents into Discovery v1, a similar API is available in v2 to upload documents into collections. You must update any workflows that you use to automate the process to account for the new arrangement of projects and collections.

If the original documents that you ingested into Discovery v1 are no longer available, you can use the query API to extract the document text from Discovery v1. You can then add the text to a collection in Discovery v2. For more information, see Recovering documents.

Crawled data

If you crawled data from an external data source in v1, you can continue to crawl data from the same external data source in v2. All of the same data sources are supported.

To use data from an external data source, you must re-create the collections within a v2 project, and configure how the data source is crawled. For more information, see Overview of data sources.

The service needs time and resources to crawl and ingest documents from external data sources. Re-create the connectors one at a time. Factor the time it takes to recrawl the data into your migration plan schedule.

Prebuilt data collections

The following built-in data source collections are not available in v2:

Watson Discovery News

This pre-enriched data source is not offered in v2. For more information about an alternative way to get news data, see Using a news service with v2.

COVID-19 kit

This pre-built collection was designed to help you fuel a dynamic chatbot that is built with IBM® watsonx™ Assistant and Discovery to answer your customers' questions about COVID-19. In v2, you can build a similar solution. Create a *Conversational Search* project type with collections that crawl trusted websites for answers to COVID-19 questions.

Ingesting data

To ingest v1 data into a Discovery v2 instance, complete the following steps:

1. Create a v2 service instance.
2. Create a project.
3. Add a collection to the project.
    - Uploaded data:

      From the API, you create a collection and add documents to it with two separate methods. Use the Create a collection method to create the collection. Next, add the same source documents that you added to your v1 collection to the v2 collection. Use the Add document or Update document methods. To assign the same v1 document ID to the document as you add it to the v2 collection, append the document ID to the endpoint. For more information, see Retaining document IDs.

      From the v2 product user interface, upload the same source documents that you added to your v1 collection to the v2 collection.

    - Crawled data: You cannot crawl data from an external data source programmatically in v2. From the product user interface, re-create the connection to the external data source, and then crawl the external data source from scratch.

4. From the product user interface, you can configure the Discovery v2 collection. For example, you can choose whether to enable optical character recognition. For an external data source, you can set the crawl schedule.

5. Apply enrichments to your data. You can apply pre-built Natural Language Processing enrichments or custom enrichments that you create.

   In v1, enrichments are associated with the configuration that is generated when you create the environment. In v2, enrichments are associated with the collection configuration. Some enrichments are applied to your collection by default, depending on the type of project used. For more information, see Default project settings. In v2, you can configure the collection to use any subset of available enrichments on the fields of your document.

Retaining document IDs

Document IDs are assigned to the documents that you add to a v2 collection when you upload them from the product user interface or add them by using the Add a document API method.

You might want to retain the IDs of your v1 documents in v2 if you are using processes that depend on these unique identifiers. For example, regression testing for the application might verify that specific documents are returned by checking the document IDs. Relevancy training uses the document IDs to track documents between training runs. These processes are easier to adapt if the document IDs are the same between your v1 and v2 instances. Otherwise, the processes that are used with the Discovery v1 instance must be remapped to the IDs that are assigned to the documents after they are added to the Discovery v2 instance.

If you specified your own documents IDs when you added documents to the v1 service instance, you can retain the IDs by using the *Update a document* method instead of the *Add a document* method. With the update method, you can assign a document ID to the document as you add it to the v2 collection. For more information, see Update a document.

Note: If your data is stored in a JSON file, an array in the original document generates a document ID with a number appended to it. For example, `original_id_n`. To retain the original document ID without the number suffix, remove the array in the JSON file. Change `[ {"name": "value"} ]` to `{"name": "value"}`, for example.

If your v1 documents have system-generated IDs, you can submit an empty search query to retrieve a list of the documents and their IDs. You can then assign the same ID to each document as you add it to your new collection in v2.

Recovering documents

In some cases, the original documents that were ingested into Discovery V1 are no longer available. You can use the Discovery v1 instance to retrieve information from the document. Discovery creates a text copy of each document that it ingests. The copy is text only, so any documents in HTML, PDF, or other nontext formats are converted to a text-only version.

Important: You can recover only the first 10,000 documents in a collection by using this method. For more information about a way to recover more than 10,000 documents, see Recovering more than 10,000 documents from a collection.

To transfer document information from v1 to v2, complete the following steps:

1. Extract the documents from v1 by using the API to submit an empty query.

   For example, `GET {url}/v1/environments/{environment_id}/collections/{collection_id}/query?q=`.

   The API returns the results. The `matching_results` field specifies the total number of results. The results object returns the matching documents. Each document is returned as a separate JSON object. It returns a maximum of 10 documents by default.

```
{
   "matching_results":  34,
   "session_token":  "nnn",
   "results":  [
      {"{result objects}":"{maximum of 10 by default}"}
   ]
}
```

2. You can use the `count` and `offset` parameters to page through the query results and save all of the documents.

   For example, to get 100 documents at a time, you can set the `count` to `100` and `offset` to `0` and submit the query.

```
GET {url}/v1/environments/{environment_id}/collections/{collection_id}/query?q=&count=100&offset=0
```

   Next, you can again set the count to 100, but this time set the offset to 100 to get the next 100 documents.

```
GET {url}/v1/environments/{environment_id}/collections/{collection_id}/query?q=&count=100&offset=100`
```

   Repeat this process, incrementing the offset by 100 until you retrieve all of the documents.

3. Prepare the exported documents to be ingested into v2.

   Each resulting JSON file that you get from Discovery v1 contains data that is extracted from the original document, such as text, html, and other fields. If custom metadata was associated with the document when it was uploaded to v1, it is also present in the JSON file. In addition, the file contains several fields that were generated by the v1 analysis. Retain only a subset of this data as part of the document that you add to Discovery v2.

   The following tips can help you decide which fields to keep:

   - Include the `text` field or any other field with textual content that you want to be able to enrich or search in Discovery v2.
   - Include any custom metadata that is stored in the document. This metatdata is typically specific to the application that uses Discovery and is used to filter documents in a search. For example, `metadata.customer_id`.
   - Do not include enrichments from Discovery v1. For example, `enriched_text.entities`. Discovery v2 generates its own enrichments.
   - Exclude fields that are generated by Discovery unless they are used by your application and contain information that is unique to the v1 version of the document. In that case, rename the field so that it does not get replaced when the document is ingested into Discovery v2. For example, `extracted_metadata.publicationdate` is a field that is generated by Discovery when a document is ingested. Maybe you want to retain the `metadata.parent_document_id` information from v1 to understand how subdocuments were originally generated from a single source document.
   - Avoid fields that have reserved field names. For more information, see How fields are handled.

4. Ingest each edited v1 JSON document into the Discovery v2 instance. The Discovery v1 document ID can be maintained in Discovery v2. For more information about how to retain the document ID, see Retaining document IDs.

Recovering more than 10,000 documents from a collection

A query can only return up to 10,000 documents. However, if you want to recover more than 10,000 documents from your collection, you need a way to separate the documents into non-overlapping subgroups. Each subgroup should contain fewer than 10,000 documents that can be returned by a query. Then, you can paginate through the results to retrieve the documents.

> **Note:** Pagination for results is restricted to the maximum of 10,000 documents that are returned by the query. Specifically, the combined use of the `count` and `offset` pagination parameters cannot exceed 10,000 documents.

One way to separate the documents into non-overlapping subgroups is to leverage a field that exists in every document and contains a unique value. For example, the SHA-1 field contains a hash of the original source file and is formatted as a hexadecimal string value. You can use the first character of the field as a way of dividing the collection into subgroups. Because SHA-1 contains a hexadecimal value, the first character can have up of 16 possible values (0-9 or a-f). If you filter by the `first_char_of (SHA-1) == 0`, it might return approximately 1/16 of the entire collection. You can then loop through each of the possible 16 values to get the rest of the documents. If optimum number of documents are not returned in one of the subgroups, you can use the first 2 characters of the SHA-1 field to divide the collection into 256 subgroups instead.

Transferring relevancy training

Relevancy training that was done in Discovery v1 can be transferred to Discovery v2. Transferring the training works best with a Discovery v2 project that has one collection that contains the same documents from the Discovery v1 collection.

Even if collections were added or documents changed, the relevancy training can be transferred. However, you must update the training to account for the changes.

To transfer relevancy training, complete the following steps:

1. Load the documents in Discovery v2.
2. Programmatically download the queries that were used for relevancy training in Discovery v1. For more information, see List training data.
3. Programmatically re-create the relevancy training data in Discovery v2. Add each training query separately by using the *Create a query* method. For more information, see Create a training query.

   Be sure to specify the v2 collection ID. You must also specify the document ID also.

   > **Note:** If you did not retain the document IDs between the v1 and v2 collections, then you must find the v2 document ID that corresponds to the v1 document ID that is referenced in the downloaded query example.

Transferring models

You can reuse some of the models that you created in v1 with your v2 project.

Smart Document Understading (SDU) models

You can import an SDU model that was built with Discovery v1 into Discovery v2. However, the performance of the model might differ between versions. Compare the results of the v1 SDU model in v2 to verify that the behavior is the same. You cannot edit the imported v1 SDU model. If the imported model can't recognize document elements that it recognized in v1 and that are important to your use case, you must re-create the SDU

model in the Discovery v2 product user interface. For more information, see Exporting SDU models in the v1 documentation and importing the SDU model in the v2 documentation.

Machine learning models

You cannot deploy models directly to Discovery v2 service instances from Knowledge Studio. Instead, you must export the machine learning models from Knowledge Studio, and then import them into Discovery. The model must have been exported from Knowledge Studio after 16 July 2020. If you have a model that was exported before that date, you must reexport the model from Knowledge Studio. Only paid Knowledge Studio plans support exporting models.

For more information, see one of the following topics:

- **IBM Cloud Pak® for Data**: Exporting a machine learning model
- **IBM Cloud**: Deploying a machine learning model to Watson Discovery

For information about how to import a model to Discovery v2, see Importing Machine Learning models.

Update your application to use the v2 API

The Watson Developer SDKs support both Discovery v1 and v2.

These instructions assume that your application is using the latest version of the v1 API (version `2019-04-30`).

When you port an application that currently uses the Discovery v1 API to use v2, you must plan how to address the following high-level differences between the two versions.

In addition to these high-level changes, review the differences at a per-method level to understand what else you might need to change. For more information, see API version comparison.

- v2 organizes data by project and collections; there is no concept of an environment. For example, compare the following requests to get a collection:

  v1 Get collection

  ```
  GET {url}/v1/environments/{environment_id}/collections/{collection_id}
  ```

  v2 Get collection

  ```
  GET {url}/v2/projects/{project_id}/collections/{collection_id}
  ```

- In v1, relevancy training runs on a single collection. In v2, relevancy training runs on a project. The project might contain many collections. If so, relevancy training is applied across all of the collections. For information about how to transfer relevancy training, see Transferring relevancy training.

  For example, compare the following requests that return the status of relevancy training:

  v1 Get collection

  ```
  GET {url}/v1/environments/{environment_id}/collections/{collection_id}
  ```

  v2 Get project

  ```
  GET {url}/v2/projects/{project_id}
  ```

- Submitting a query is similar between the two versions. In v2, you can query all of the collections in a project or you can limit the query to one or more collections by specifying a `collection_ids` parameter. For example, compare the following requests to query data:

  v1 Query request

  ```
  POST {url}/v1/environments/{environment_id}/collections/{collection_id}/query
  ```

  Data that is submitted with the request:

  ```
  {
    "query":  "text:IBM"
  }
  ```

  v2 Query request

```
POST {url}/v2/projects/{project_id}/query
```

Data that is submitted with the request:

```
{
   "collection_ids":  [
      "{collection_id_1}",
      "{collection_id_2}"
   ],
   "query":  "text:IBM"
}
```

You can optionally omit the `collection_ids` parameter to query across all of the collections in the project.

- The `passage` parameter for a query has a new `per_document` option that ranks the documents by document quality, and then returns the highest-ranked passages per document in a `document_passages` field for each document entry in the results list of the response. If false, ranks the passages from all of the documents by passage quality regardless of the document quality and returns them in a separate passages field in the response.

- When passages are returned for a query, you can also enable answer finding. When true, answer objects are returned as part of each passage in the query results. When `find_answers` and `per_document` are both set to true, the document search results and the passage search results within each document are reordered by using the answer confidences. The goal of this reordering is to place the best answer as the first answer of the first passage of the first document. Similarly, if the `find_answers` parameter is set to true and per_document parameter is set to false, then the passage search results are reordered in decreasing order of the highest confidence answer for each document and passage.

- Both v1 and v2 support custom stop words. However, there are a few differences in how custom stop words are used:

  - There is no default custom stop words list for Japanese collections in v2.
  - When you define custom stop words in v1, your stop words list replaces the existing stop words list. In v2, your list augments the default list. You cannot replace the list, which means you cannot remove stop words that are part of the default list in v2.

Update how your application handles query results

The way that your application shows query results might need to be updated due to the following differences between the query results document syntax between the v1 and v2 queries:

- At the entity enrichment level, the following information is not supported in v2:

  - Disambiguation
  - Emotion
  - Sentiment

  The *Part of Speech* enrichment is applied automatically to documents in most project types in v2, but the index fields that are generated by the enrichment are not displayed in the JSON representation of the document.

Figure 1. Entities data structure differences

- Instead of the `count` and `relevance` in v1, v2 includes the mentions.

  Each entry in the mention corresponds to an occurrence of the entity in the document text. In the following example, seven occurrences are found. For each occurrence, a confidence score and the offsets of the mention text are displayed. You can use the offsets to highlight the mention in the document text when the result is displayed in a user interface.

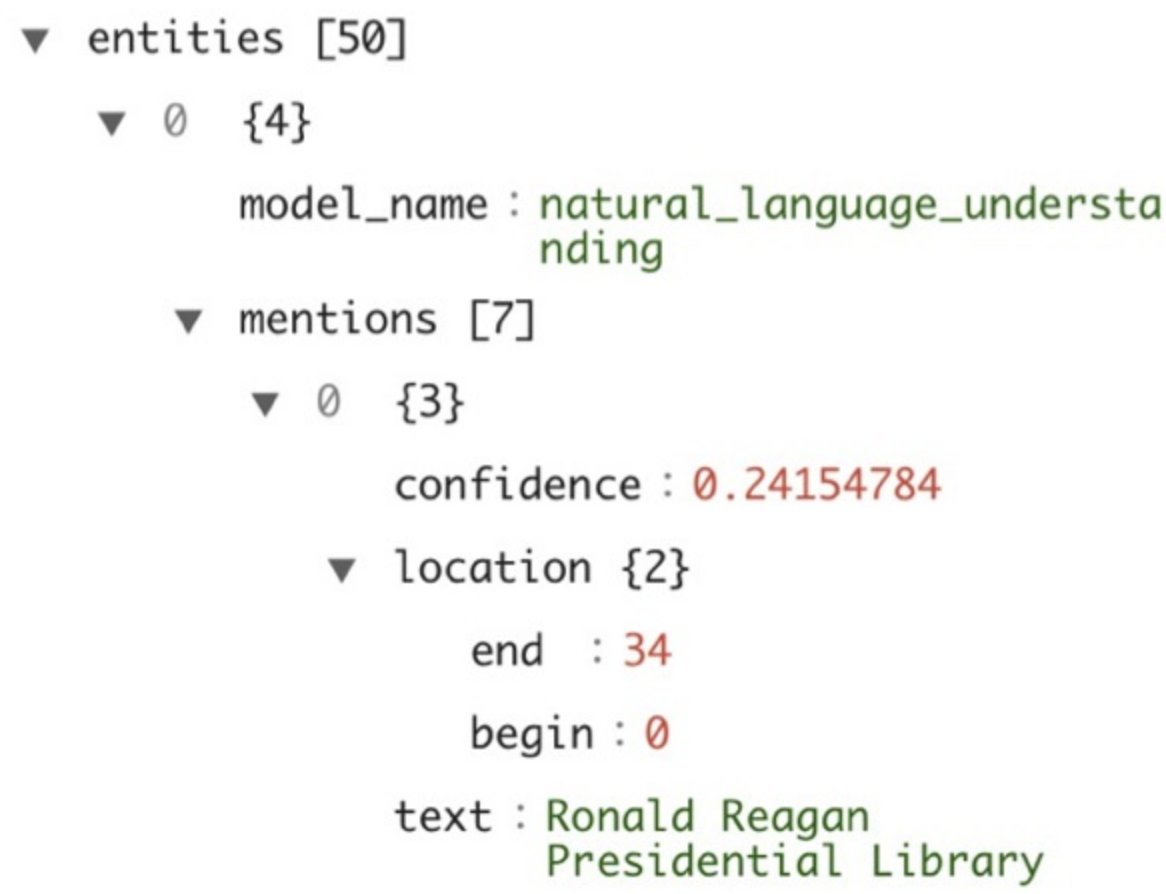

Figure 2. Entity mentions in Discovery v2

- The JSON structure of query responses is rearranged slightly in v2.
- Deduplication information is not included in the v2 query response.
- In v2, `enriched_text` is an array instead of an object.
- In Discovery v2, the Entities v2 enrichment is used. Entity type names in v2 are specified in headline case, instead of all uppercase letters. If you use a query or aggregation that specifies an entity name, you must change the capitalization. For example, change `PERSON` to `Person`.
- Fields from JSON files that are added to a collection are converted differently during ingestion between v1 and v2. If your application manipulates these results, you might need to make adjustments.

> **Note:** You can specify the `normalizations` and `conversions` objects in the Update a collection method of the API to move or merge JSON fields.

Original JSON field content	v1 representation	v2 representation	Notes
`"field": null`	`"field": null`	N/A	v1 retains the null value. v2 skips the null field altogether.
`"field": ""`	`"field": ""`	N/A	v1 retains the empty text value. v2 skips the empty text field altogether.
`"field": "value2"`	`"field": "value2"`	`"field": "value2"`	No difference.
`"field": []`	`"field": []`	N/A	v1 retains the empty array. v2 skips the field with the empty array altogether.
`"field": [ "value4" ]`	`"field": [ "value4" ]`	`"field": "value4"`	v1 retains the singleton array. v2 converts the singleton array into the value only; it is not stored as part of an array.
`"field": [ 1, 2, 3 ]`	`"field": [ 1, 2, 3 ]`	`"field": [ 1, 2, 3 ]`	No difference.
`"field": [ "v6", "v7", "v8" ]`	`"field": [ "v6", "v7", "v8"]`	`"field": [ "v6", "v7", "v8"]`	No difference.

How JSON source fields are handled

Verifying that your data was migrated successfully

To verify that the migration was successful, compare the following metrics to the metrics that you noted before the migration.

- Number of collections

  Be sure to re-create all of the collections that you used in v1 and want to keep. With the v2 List collections API method, you can get a list of collections, but you must submit a request per project. You cannot use one call to get the total number of collections per service instance.

- Number of documents per collection

  For collections with uploaded data, check the number of documents in the collection by sending an empty query with the Query a project API method. Specify the collection ID parameter to limit the results to only documents in one collection. An empty query returns all documents. Therefore, you can get the total number of documents from the `matching_results` value in the response.

  The number of documents per collection should be close to the number of documents that were stored in the same collection in v1. The numbers might not be the same.

  For crawled data, do not be surprised if the v2 collection has fewer documents. The v1 connectors do not delete documents from a Discovery collection that are deleted from the external data source. Your v2 version of the collection has a fresher crawl of the data as it exists in the external data source today.

Tip: Do not expect the search results to be the same for queries that you submit in the v1 and v2 instances.

Using a news service with v2

If you used the Watson Discovery News data source in v1 and want to create a data source with equivalent function in v2, find a news and events data provider service. Look for a service that offers a News API that extracts news articles in JSON format. You can then upload the JSON files to create a News collection in your v2 project.

Delete your v1 service instance

After your data is migrated and your applications are updated to use the new v2 service instance, be sure to delete your v1 service instance. You are charged for the v1 service instance until you delete it. For more information, see Deleting a managed service instance.

API version comparison

For most API methods, the request parameters and response bodies differ between v1 and v2. Learn about the equivalent or alternative v2 methods that you can use to do actions that are supported by the v1 API.

The comparison information assumes you are using the latest version of the v1 API (version `2019-04-30`) and compares it to the latest version of the v2 API (version `2020-08-30`).

Environments

There is no concept of an *environment* in v2. The deployment details such as size and index capacity are managed based on the service plan type. In v2, collections are organized in projects. You can create different types of projects to apply default configuration settings to the collections that you add to the projects.

There are no equivalent methods in v2 for the v1 environment methods. However, the following table shows v2 methods that serve similar functions to the corresponding v1 methods. The supported parameters and response bodies that are returned for each method differ also.

Action	v1 API	Related v2 API
Create an environment	`POST /v1/environments`	`POST /v2/projects`
List environments	`GET /v1/environments`	`GET /v2/projects`
Get environment info	`GET /v1/environments/{environment_id}`	`GET /v2/projects/{project_id}`
Update an environment	`PUT /v1/environments/{environment_id}`	`POST /v2/projects/{project_id}` v2 uses **POST** instead of **PUT**.
Delete an environment	`DELETE /v1/environment/{environment_id}`	`DELETE /v2/projects/{project_id}`
List fields across collections	`GET /v1/environments/{environment_id}/fields`	`GET /v2/projects/{project_id}/fields`

Environment API action support details

Configurations

The v2 API does not have an endpoint that is dedicated to configurations. Instead, configuration settings for projects, collections, and queries are specified directly in the API for those objects. Not all of the configuration parameters that are available in v1 are available or applicable in v2.

In the v1 configuration API, the JSON object that is used to specify a configuration object contains several parameters that are either available in different formats from other v2 endpoints or are not available in v2. The following table describes how to find related parameters in v2.

You cannot customize the conversion of documents during the ingestion process in v2 as you can in v1.

v1 configuration parameter	v2 API
`"conversions.html": { ... }`	Not available
`"conversions.image_text_recognition": { ... }`	Not available from the API. However, you can enable optical character recognition (OCR) for a collection from the product user interface to extract text from images. OCR has other benefits, too. For example, if a page in a document can't be processed, OCR converts the page into an image and scans it to ensure that the document is uploaded successfully.
`"conversions.json_normalizations": { ... }`	Moved to the Collections API.
`"conversions.pdf": { ... }`	Not available. If you used special parameters to extract text from images in PDFs, enable optical character recognition (OCR) from the product user interface for the collection that contains the PDFs instead.
`"conversions.segment": { ... }`	Not available programmatically. You can split a document at each occurrence of an SDU-generated field such as `subtitle` from the product user interface. The `segment_metadata` object with `parent_id`, `id`, and `total_segments` information is not available in v2. You can use the `metadata.parent_document_id` field to find the common parent for many document segments.
`"conversions.word": { ... }`	Not available
`"enrichments": { ... }`	`/v2/projects/{project_id}/enrichments`, `/v2/projects/{project_id}/collections/{collection_id}` Use the enrichments API to explore existing enrichments. Use the collections API to see and change the enrichments that are enabled on a field in a collection. Some enrichments are applied to the service by default based on the type of project that you create. For more details, see Default project settings. The version of the Entities enrichment that is available in v2 doesn't include the `disambiguation` field, which in v1 contains the disambiguation information for the entity and includes the entity subtype information. The following enrichments are not available in v2: • Categories • Concepts • Emotion • Relations • Semantic roles • Sentiment of Entities • Sentiment of Keywords
`"normalizations": [ ... ]`	Moved to the Collections API.
`"source": { ... }`	Not available. Configure connections to external data sources through the user interface. For more information, see Creating collections.

Configuration setting details

Collections

Action	v1 API	v2 API

Create a collection	`POST /v1/environments/{environment_id}/collections`	`POST /v2/projects/{project_id}/collections` The supported parameters and responses differ between the tv versions. See the collection notes.
List collections	`GET /v1/environments/{environment_id}/collections`	`GET /v2/projects/{project_id}/collections` In v2, only the collection ID and name of each collection are re the list. You must use the *Get collection* method to return more about each collection.
Get collection details	`GET /v1/environments/{environment_id}/collections/{collection_id}`	`GET /v2/projects/{project_id}/collections/{collect` See the collection notes.
Update a collection	`PUT /v1/environments/{environment_id}/collections/{collection_id}`	`POST /v2/projects/{project_id}/collections/{collect`
Delete a collection	`DELETE /v1/environments/{environment_id}/collections/{collection_id}`	`DELETE /v2/projects/{project_id}/collections/{collect` In v2, the `status` field is not returned in the response.
List collection fields	GET /v1/environments/{environment_id}/collections/{collection_id}/fields v1 lists the fields per collection.	GET /v2/projects/{project_id}/fields v2 lists fields per project instead. You can pass a single collecti with the `collection_ids` parameter to get fields from a sing collection.

Collections API support details

Collections API notes

The following table shows the important differences between the v1 and v2 collection APIs.

Method	Notes
Create a collection	The v2 response doesn't include the `status` and `configuration_id` fields. You can get status information for a specific document by using the Get document details method. The objects `disk_usage`, `training_status`, and `crawl_status` are not present in the response body in v2. The `document_counts` object is not present in the response body in v2 currently. Training status is returned in the Get project method response. The other information is not available in v2. In v2, you can define the enrichments to apply to the documents in the collection by specifying an optional `enrichments` object.
Get collection details	The v2 response doesn't include the `status` and `configuration_id` fields. You can get status information for a specific document by using the Get document details method. The objects `document_counts`, `disk_usage`, `training_status`, and `crawl_status` are not present in the response body in v2. Training status is returned in the *Get project* method response. The other information is not available in v2. For example, you cannot get the document count for a collection and cannot get the crawl status for a collection that connects to an external data source in v2. In v2, you can get information about the enrichments that are applied to the collection.
Update a collection	v2 uses `POST` instead of `PUT`. In v2, you can update the enrichments that are applied to the documents in the collection by specifying an optional `enrichments` object. The v2 response doesn't include the `status` and `configuration_id` fields.

Collections API notes

Query modifications

The method that was available in v1 for configuring tokenization programmatically is not supported in the v2 API.

v1 API	v2 API
Tokenization dictionary API	Not available.
Expansions v1 API	Expansions v2 API
Stopwords v1 API	Stopwords v2 API

Query modifications API support details

Documents

Action	v1 API	v2 API
List documents	Not available from the v1 API	GET /v2/projects/{project_id}/collections/{collec
Create a document	POST /v1/environments/{environment_id}/collections/{collection_id}/documents	POST /v2/projects/{project_id}/colle Unlike v1, the v2 response does not include a not by using the *Get document details* method in v2.
Update a document	POST /v1/environments/{environment_id}/collections/{collection_id}/documents/{document_id}	POST /v2/projects/{project_id}/collection When you update a document that was split, all c
Get document details	GET /v1/environments/{environment_id}/collections/{collection_id}/documents/{document_id}	GET /v2/projects/{project_id}/collections/{collec In v2, there is no `statusDescription`. v2 has that are associated with the child documents tha
Delete a document	DELETE /v1/environments/{environment_id}/collections/{collection_id}/documents/{document_id}	DELETE /v2/projects/{project_id}/collection Segments of an uploaded document cannot be d request that includes the parent_document_id o

Documents API support details

v2 introduces a custom header that is named `X-Watson-Discovery-Force` that is not available in v1. You must include the header when you perform an operation on data that is shared across many collections to indicate that you want to perform the operation in each collection. If you do not include the header, a `403` error is returned.

Fields from JSON files that are added to a collection are converted differently during ingestion between v1 and v2. For more information about how JSON files are stored in the v2 index, see JSON files.

Queries

Action	v1 API	v2 API
Query a collection	Supports a GET or POST request. GET or POST /v1/environments/{environment_id}/collections/{collection_id}/query	Queries a project. To specify a single collection, include the {collection_id} parameter. Supports a POST request only. POST /v2/projects/{project_id}/query
Query multiple collections	GET or POST /v1/environments/{environment_id}/query	POST /v2/projects/{project_id}/query
Query system notices	GET /v1/environments/{environment_id}/collections/{collection_id}/notices	GET /v2/projects/{project_id}/collections/{collection_id}/notices
Query multiple collection system notices	GET /v1/environments/{environment_id}/notices	GET /v2/projects/{project_id}/notices
Get Autocomplete suggestions	/v1/environments/{environment_id}/collections/{collection_id}/autocompletion	GET /v2/projects/{project_id}/autocompletion See the query notes.

Documents API support details

Some query result configurations are applied to the service by default based on the type of project that you create. For more details, see Default project settings.

Query notes

- v2 queries return results from all of the collections in the project. To restrict the query to use only certain collections within the project, use the `collection_ids` query parameter. You cannot query multiple collections that are added to different projects with one v2 query request.
- v2 results include a `confidence` field, but not a `score` field.

  The confidence score replaced the score information in v1, but score was retained for backward compatibility. In v2, only the confidence field is returned.
- Use POST calls (instead of GET calls) to submit queries with v2.
- v1 queries accept many parameters. The *Query parameters comparison* table maps v1 parameters to v2 parameters.

v1 parameter	v2 parameter	Notes
N/A	collection_ids	Use this parameter in v2 to specify collection ids.
filter	filter	Same expression language.
query	query	Same expression language.
natural_language_query	natural_language_query	No notes.
passages	passages	The passage format changed and was enhanced in v2. The `passages:true` parameter changed to `passages.enable:true`. In addition to the `count`, `characters`, and `fields` options, you can specify `per_document`, which ranks the documents by document quality, and then returns the highest-ranked passages per document. You can also specify `find_answers` to return an answer object per passage, which contains a succinct answer to the query.
aggregation	aggregation	Same expression language.
count	count	No notes.
offset	offset	No notes.
return	return	No notes.
sort	sort	No notes.
highlight	highlight	If `passages.enabled` and `passages.per_document` are `true`, then passages are returned for each document instead of highlights.
spelling_suggestions	spelling_suggestions	No notes.
deduplicate	**N/A**	Not supported in v2.
similar	similar	The format changed in v2. The `similar:true` parameter changed to `similar.enable:true`. The `document_ids` and `fields` parameters changed from strings to string arrays. The `document_ids` parameter now is required if `enabled` is true.
bias	**N/A**	Not supported in v2.

Query parameters comparison

Training data

You can use the v1 training data API to work with two related objects:

- trained queries
- examples that are used to train the queries

These two objects have separate API endpoints in v1. In v2, the examples that are used to train each query are provided together with the query and only one endpoint is used to work with the training data.

For example, to add a trained query and its training example documents in v2, you use the request `POST /v2/projects/{project_id}/training_data/queries` and pass the query and all examples in the payload of one call. Similarly, if you want to update one example in the training set in v2, you must pass the query and the modified example (along with all of the other examples) to the v2 update endpoint. In v1, to update the example information, you use the update example endpoint to modify one example only.

Another important difference between v1 and v2 is that in v1, the trained model is associated with a particular collection. In v2, the trained model is associated with a project. You can use the data from multiple collections within a project to train a relevancy model. When you create or update training examples in v2, the API requires the `collection_id` for the collection where the document is stored.

Action	v1 API
List training data	`GET /v1/environments/{environment_id}/collections/{collection_id}/training_data`
Add query to training data	`POST /v1/environments/{environment_id}/collections/{collection_id}/training_data`
Delete all training data	`DELETE /v1/environments/{environment_id}/collections/{collection_id}/training_data`
Get details about a query	`GET /v1/environments/{environment_id}/collections/{collection_id}/training_data/{query_id}`
Delete a training data query	`DELETE /v1/environments/{environment_id}/collections/{collection_id}/training_data/{query_id}`
List examples for a training data query	`GET /v1/environments/{environment_id}/collections/{collection_id}/training_data/{query_id}/examples`
Add example to training data query	`POST /v1/environments/{environment_id}/collections/{collection_id}/training_data/{query_id}/examples`
Delete example for training data query	`DELETE /v1/environments/{environment_id}/collections/{collection_id}/training_data/{query_id}/examples/{example_id}`
Change label or cross-reference for example	`PUT /v1/environments/{environment_id}/collections/{collection_id}/training_data/{query_id}/examples/{example_id}`

Get details for a training data example	`GET /v1/environments/{environment_id}/collections/{collection_id}/training_data/{query_id}/examples/{example_id}`

Training data API support details

User data

The user data API is the same in v2 and v1.

Action	v1 API	v2 API
Delete	`DELETE /v1/user_data`	`DELETE /v2/user_data` Similar to v1. Use `customer_id` to delete the data associated with that customer ID.

User data API support details

Events and feedback

The v1 events and feedback API (`/v1/events`) is not available in v2.

Credentials

The v1 credentials API (`/v1/environments/{environment_id}/credentials`) is not available in v2. The function is available from the v2 product user interface.

Gateway configuration

The v1 gateways API (`/v1/environments/{environment_id}/gateways`) is not available in v2. The function is available from the v2 product user interface. For more information, see Installing IBM Secure Gateway for on-premises data.

Status codes

For almost every API method, the status codes that are returned for v2 requests are different from the status codes that are returned for v1 requests.

Migration FAQ

Find answers to questions that are commonly asked about migrating from Discovery v1 to v2.

Do the two versions have all the same features?

There are many feature differences between the two versions. For a full feature comparison, see Getting the most from Discovery.

How do I know which version I'm using now?

When you open the product user interface in v2, the following page is displayed:

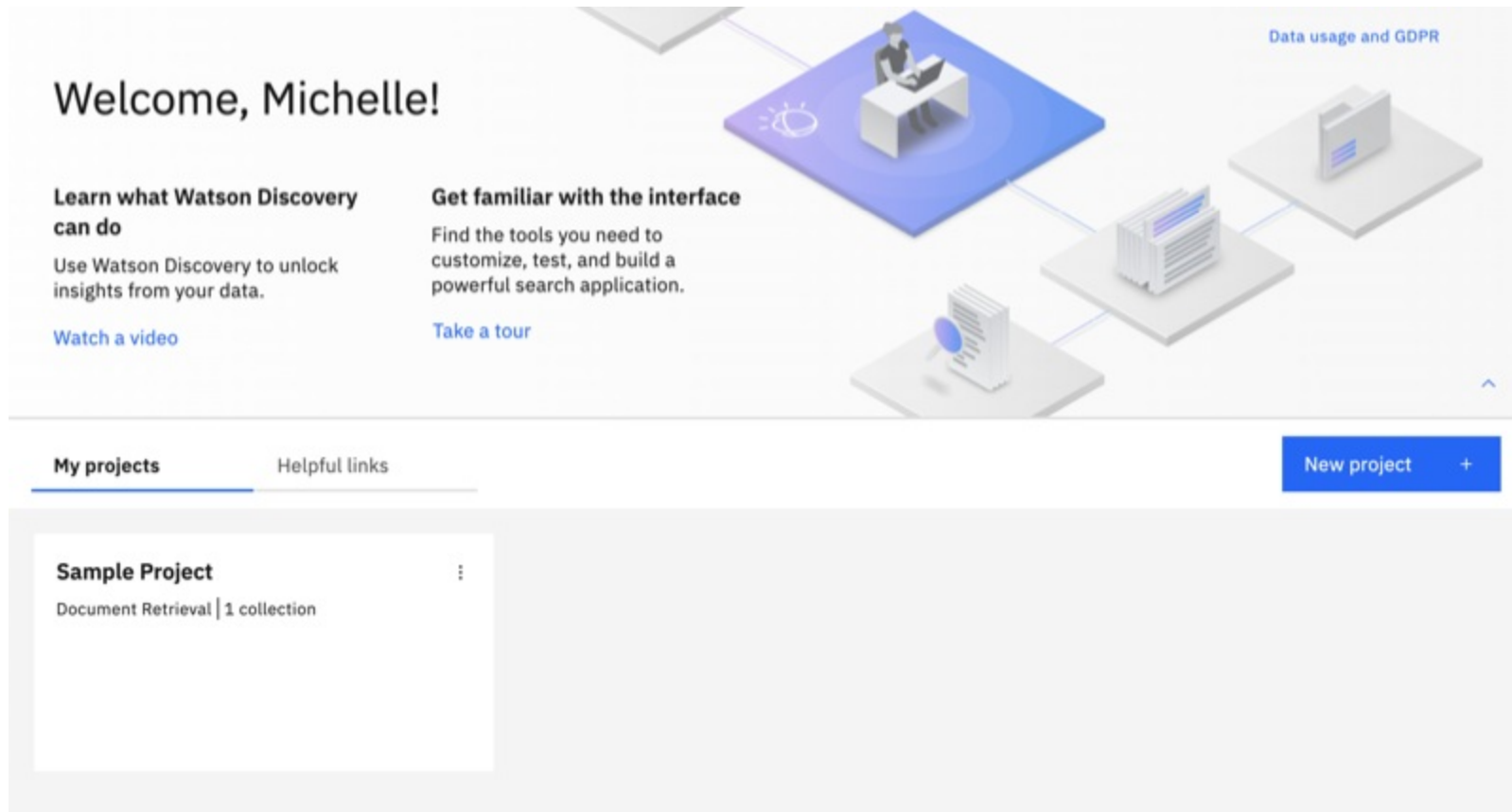

Figure 1. Home page from the Sample Project

How long will the migration take?

The time you need to set aside for the migration differs based on the amount of data you want to retain in your existing v1 service instance.

Do I need to update my existing applications for them to work with v2?

Yes. You will need to edit any existing applications to account for changes that are introduced with Discovery v2. For more information, see the API version comparison.

To get started, see Migrating to Discovery v2.

Migrating enrichments from Watson Explorer

If you have resources from IBM Watson Explorer, some of them can be migrated to IBM Watson® Discovery.

Types of resources that can be migrated

The following types of resources can be migrated from Watson Explorer to Discovery:

- From Watson Explorer Analytical Components: User dictionaries
- From Watson Explorer oneWEX: Dictionaries, character patterns, and facets.

To analyze data with these migrated enrichments, you can use a Content Mining project. The tools in the associated Content Mining application are similar to tools that are available in Watson Explorer.

- For more information about how to create a Content Mining project, see Creating projects.
- For more information about how to apply enrichments to a collection in the Content Mining application, see Applying the annotator.

Importing dictionaries from Watson Explorer Analytical Components

You can import user dictionaries from IBM Watson Explorer Analytical Components.

Tip: The default file location and name for dictionaries that are saved in Watson Explorer Analytical Components is `${primary_server_node}/{primary_configuration}/{collection_ID}/{dictionary_name}.fdic.xml`.

1. Download your user dictionaries from Watson Explorer Analytical Components.
2. From your Discovery Content mining project, open the Content Mining application.
3. From the analysis view of your collection, click the **Collections** link in the breadcrumb to open the *Create a collection for your analytics solutions* page of the Content Mining application.
4. To create an annotator, click **collection**, and then select **custom annotator** from the list.

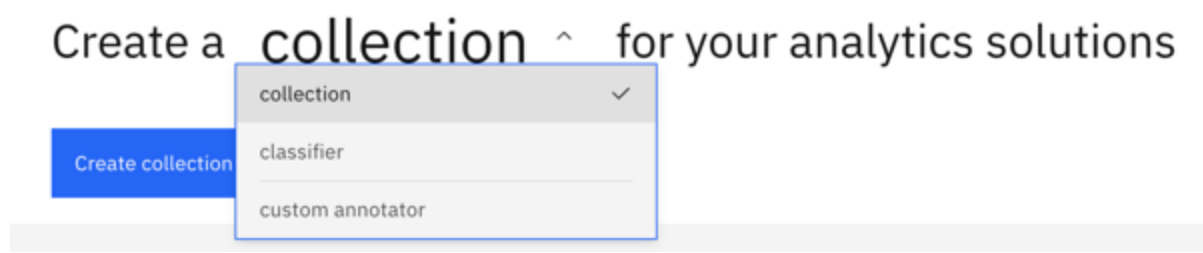

Figure 1. Collection menu

5. Click **Create custom annotator**.
6. Name your annotator, and then optionally add a description.
7. From the **Annotator Type** menu, select **Dictionary**, and then click **Next**.
8. Click the **Import** button, and then select the `{name}.fdic.xml` dictionary file that you want to import.
9. Click **Save**.

Uploading dictionaries from Watson Explorer oneWEX

You can import dictionaries from IBM Watson Explorer oneWEX.

1. From Watson Explorer oneWEX, Version 12.0.0 or later modifications or fix packs, download the dictionary CSV file.
   - Log in to the oneWEX administrator console.
   - Open the **Resource** tab.
   - Select dictionary enrichments, open the dictionary tab, and click the download icon. The dictionary is downloaded as a CSV file.
2. From your Discovery Content mining project, open the Content Mining application.
3. From the analysis view of your collection, click the **Collections** link in the breadcrumb to open the *Create a collection for your analytics solutions* page of the Content Mining application.
4. To create an annotator, click **collection**, and then select **custom annotator** from the list.

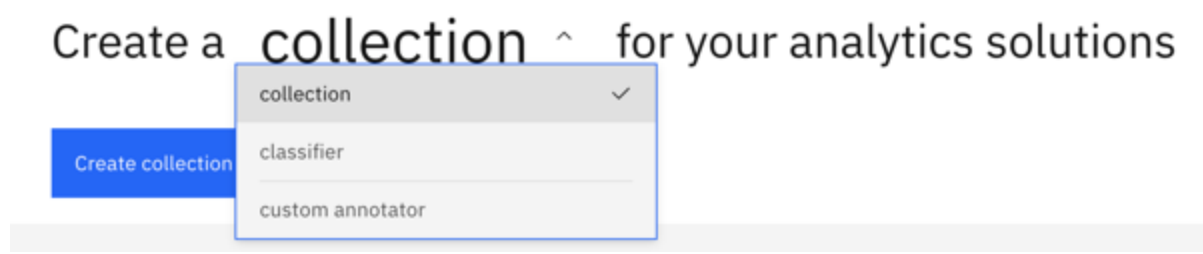

Figure 2. Collection menu

5. Click **Create custom annotator**.
6. Name your annotator, and then optionally add a description.
7. From the **Annotator Type** menu, select **Dictionary**, and then click **Next**.
8. Click the **Import** button, and then upload the CSV file of the dictionary that was downloaded from oneWEX.
9. Click **Import**, and then click **Save**.

Importing character patterns from Watson Explorer oneWEX

You can import character patterns from IBM Watson Explorer oneWEX.

1. From your Discovery Content mining project, open the Content Mining application.
2. From the analysis view of your collection, click the **Collections** link in the breadcrumb to open the *Create a collection for your analytics solutions* page of the Content Mining application.
3. To create an annotator, click **collection**, and then select **custom annotator** from the list.

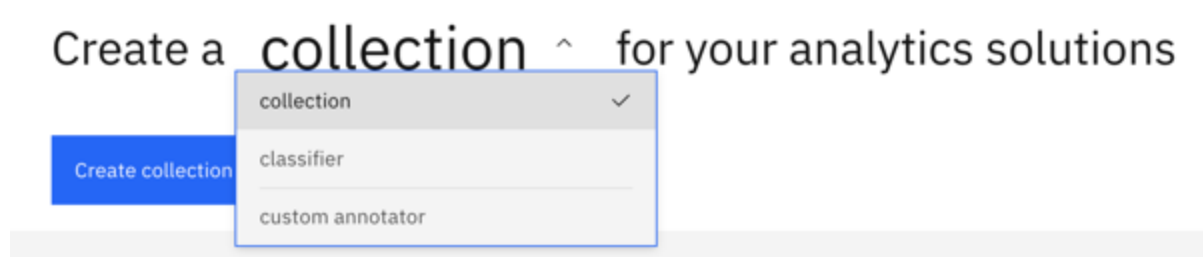

Figure 3. Collection menu

4. Click **Create custom annotator**.
5. Name your annotator, and then optionally add a description.
6. From the **Annotator Type** menu, select **Regular expression**, and then click **Next**.
7. Click the **Import** button.
8. Select the JSON file that you want to import, and then click **Save**.

Importing facets from Watson Explorer Content Analytics Studio IBM Cloud Pak for Data

> **Note:** You can import a PEAR file to use as the machine learning source file from IBM Cloud Pak for Data deployments only.

You can show Content Analytics Studio facets in the Content Mining application. Only facets with a UIMA Feature of type `Literal Value` are displayed.

For more information about how to import Content Analytics Studio machine learning models for use in other project types, see Use imported ML models to find custom terms.

1. From the Watson Explorer Content Analytics Studio, export the machine learning model that defines the facets that you want to use. The model file must have a `.pear` extension.
2. In the export configuration, remove the facet path, but keep the subfacet value. Set the Index Field name to the Facet Tree Path in Content Analytics Studio.

   For more information, see Creating Custom PEAR Files for Use with Lexical Analysis Streams .
3. From your Discovery Content mining project, open the Content Mining application.
4. From the analysis view of your collection, click the **Collections** link in the breadcrumb to open the *Create a collection for your analytics solutions* page of the Content Mining application.
5. To create an annotator, click **collection**, and then select **custom annotator** from the list.

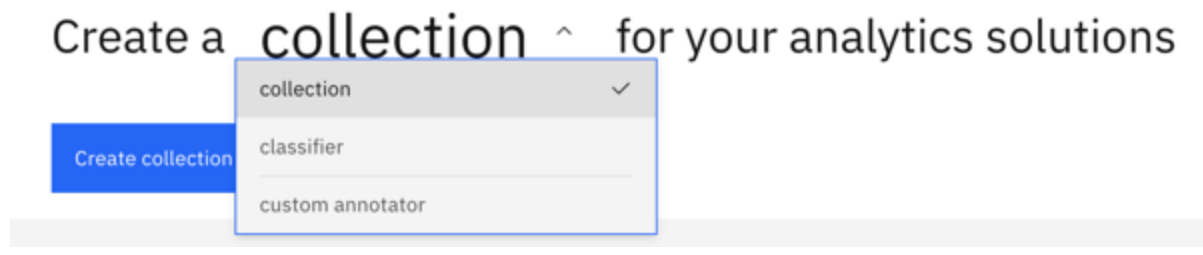

Figure 4. Collection menu

6. Click **Create custom annotator**.
7. Name your annotator, and then optionally add a description.
8. From the **Annotator Type** menu, select **PEAR File**, and then click **Next**.
9. Click **Select file** to find the .pear file that you exported.
10. Specify a facet path, and then click **Save**.

Migrating Knowledge Studio solutions

Use custom models and other resources that you created in Knowledge Studio by migrating them to Discovery.

Using a model as is

To start using your Knowledge Studio model immediately, export the model from Knowledge Studio and import it to Discovery as a machine learning enrichment.

When you import a Knowledge Studio model to use as is in Discovery, root-level entity types that were defined in the model can be recognized when they occur in your documents. Any mentions of entity subtypes that occur are identified as mentions of the parent entity type. The subtype entities themselves are not preserved. If you want the model to continue to distinguish between different subtypes of an entity, you must take extra steps. For more information, see Retaining subtype information.

> **Note:** You cannot continue to update a model that you import as an ML enrichment.

The following types of models can be imported and used as is:

- Rule-based models created in Knowledge Studio that find entities in documents based on rules that you define. (File format: .pear)

- Machine learning models created in Knowledge Studio that understand the linguistic nuances, meaning, and relationships specific to your industry (file format: .zip)

The models that you can add depend on your deployment type:

- IBM Cloud You can add models that were created with a IBM Watson® Knowledge Studio instance that is hosted in IBM Cloud only.
- IBM Cloud Pak for Data You can add models that were created with an instance of IBM Watson® Knowledge Studio that is hosted on IBM Cloud Pak® for Data or IBM Cloud.

For more information, see Using imported ML models to find custom terms .

Using a corpus as training data

Discovery has an entity extractor tool that you can use to define a type system. The entity extractor user interface is similar to the Knowledge Studio user interface that is used to annotate documents that you add to corpus for a machine learning model. However, in Knowledge Studio, you define root-level entities only, not subtypes or relationships.

As an alternative to importing a Knowledge Studio model as is and applying it as an enrichment, you can also import a Knowledge Studio corpus. When you add a Knowledge Studio corpus to the Discovery entity extractor tool, any root-level entities from the corpus are represented as new entities in the Discovery entity extractor workspace. Entity subtypes are not recognized. Although, you can take extra steps to retain subtype information.

Relations and coreferences from the Knowledge Studio machine learning model are not represented, neither are any custom dictionaries that are associated with the model.

Things to consider when choosing whether to import a model or import a corpus:

- You can continue to edit the type system when you import the corpus. When you import a trained model, you cannot subsequently edit it in Discovery.
- An imported model that you apply to a collection as an enrichment can recognize any entity subtype, relation, and coreference information that the original model was trained to recognize in addition to root-level entities. An entity extractor enrichment can find and tag entities only.

For more information, see Importing a Knowledge Studio corpus.

Retaining subtype information

When you import a Knowledge Studio model to Discovery, any subtypes that were defined in the model are identified as mentions of the parent entity type. The subtype entities themselves are not preserved. To retain the subtype information, you must *flatten* your type system by converting entity subtypes into new root-level entity types.

Follow these steps only if you are sure that the subtype distinctions add significant value to the model. In many use cases, using the root-level entity types is sufficient.

⚠ **Important:** You cannot use this procedure to retain subtypes if any of the documents in your corpus were pre-annotated with the Natural Language Understanding service. Make sure that your flattened type system doesn't surpass the allowed number of entity types for your plan. For more information, see Entity extractor limits .

For example, your model might have entity types with the following hierarchy:

```
APPLIANCES
FURNITURE
  PATIO
  LIVING
  DINING
```

A flattened version of the type sytem looks like this:

```
APPLIANCES
FURNITURE_NONE
FURNITURE_PATIO
FURNITURE_LIVING
FURNITURE_DINING
```

A useful approach for flattening the type system involves the following changes:

- Add the parent entity type label (`FURNITURE`) as a prefix to the label of each child subtype to produce a new root-level entity that preserves the hierarchical relationship in its label. For example, `FURNITURE_PATIO`, `FURNITURE_LIVING`, and `FURNITURE_DINING`.
- Append the word *NONE* to the parent root-level entity label to identify it as the parent. For example, `FURNITURE_NONE`.
- Leave the labels of entity types that don't have subtypes unchanged. For example, the label `APPLIANCES` doesn't change.

To retain entity subtype information, complete the following steps:

1. Ensure that the annotation and training of the Knowledge Studio model is completed and the model is ready to be deployed.

2. Export the type system that was used to annotate the documents in your corpus from Knowledge Studio as a .json file.

   Follow the appropriate steps for exporting based on your Knowledge Studio deployment type:

   - IBM Cloud Uploading resources from another workspace
   - IBM Cloud Pak for Data Uploading resources from another workspace

3. Modify the type system JSON file. For each subtype, add a new root-level entity type.

   For example, the original type system might contain the following types:


```
{
   "id":"b9d6caa2-90ac-47ff-91f6-2149b8ffcf20",
   "label":"FURNITURE",
   "sireProp":{
      "mentionType":null,
      "subtypes":["PATIO","LIVING","DINING"],
      "roles":["b9d6caa2-90ac-47ff-91f6-2149b8ffcf20","93ba1f27-173f-4714-b31e-77bdd8cb9932"],
      "clazz":null,
      "color":"black",
      "hotkey":"m",
      "backGroundColor":"#00FFFF",
      "active":true,
      "roleOnly":false},
      "creationDate":1610611788484,
      "source":null,
      "modifiedDate":0,
      "typeType":null,
      "typeClass":null,
      "typeVersion":null,
      "typeDesc":null,
      "typeSuperType":null,
      "typeSuperTypeId":null,
      "typeCreateDate":null,
      "typeUpdateDate":null,
      "typeProvenance":null,
      "alchemyAPITypes":null,
      "nluAPITypes":null},
```


To convert the subtypes to new root-level types, make the following change:


```
{
   "id":"b9d6caa2-90ac-47ff-91f6-2149b8ffcf20",
   "label":"FURNITURE_NONE",
   "sireProp":{
      "mentionType":null,
      "subtypes":null,
      "roles":["b9d6caa2-90ac-47ff-91f6-2149b8ffcf20","93ba1f27-173f-4714-b31e-77bdd8cb9932"],
      "clazz":null,
      "and so on"
      }
},
{
   "id":"b9d6caa2-90ac-47ff-91f6-2149b8ffcf20",
   "label":"FURNITURE_PATIO",
   "sireProp":{
      "mentionType":null,
      "subtypes":null,
      "roles":["b9d6caa2-90ac-47ff-91f6-2149b8ffcf20","93ba1f27-173f-4714-b31e-77bdd8cb9932"],
      "clazz":null,
      "and so on"
      }
},
{
   "id":"b9d6caa2-90ac-47ff-91f6-2149b8ffcf20",
   "label":"FURNITURE_LIVING",
   "sireProp":{
      "mentionType":null,
      "subtypes":null,
      "roles":["b9d6caa2-90ac-47ff-91f6-2149b8ffcf20","93ba1f27-173f-4714-b31e-77bdd8cb9932"],
      "clazz":null,
      "and so on"
      }
},
{
   "id":"b9d6caa2-90ac-47ff-91f6-2149b8ffcf20",
   "label":"FURNITURE_DINING",
```

```
    "sireProp":{
        "mentionType":null,
        "subtypes":null,
        "roles":["b9d6caa2-90ac-47ff-91f6-2149b8ffcf20","93ba1f27-173f-4714-b31e-77bdd8cb9932"],
        "clazz":null,
        "and so on"
        }
},
```

4. Assign a unique ID to each new root-level entity type.
5. Export the corpus for your machine learning model from Knowledge Studio as a compressed file.

   Follow the appropriate steps for exporting based on your Knowledge Studio deployment type:

   - IBM Cloud Uploading resources from another workspace
   - IBM Cloud Pak for Data Uploading resources from another workspace
6. In the downloaded corpus, for all mentions with a subtype defined, update the type information for the mention to specify the new root-level entity type.

   For example, the original type system might include the `PATIO` subtype mention:

```
{
   "id" : "Blogs_shopper.com_dc5cf4764d91f87575b17ac8a5268462.en-M92",
   "source" : "IMPORT",
   "properties" : {
      "SIRE_ENTITY_CLASS" : "SPC",
      "SIRE_MENTION_CLASS" : "SPC",
      "SIRE_ENTITY_LEVEL" : "NONE",
      "SIRE_ENTITY_SUBTYPE" : "PATIO",
      "SIRE_MENTION_ROLE" : "FURNITURE",
      "SIRE_MENTION_TYPE" : "NONE"
   },
   "type" : "FURNITURE",
   "begin" : 3221,
   "end" : 3234,
   "inCoref" : false
},
```

   Replace the value of the `SIRE_MENTION_ROLE` and `type` for the mention with the new root-level entity label, such as `FURNITURE_PATIO`. Specify `NONE` as the `SIRE_ENTITY_SUBTYPE` value.

```
{
   "id" : "Blogs_shopper.com_dc5cf4764d91f87575b17ac8a5268462.en-M92",
   "source" : "IMPORT",
   "properties" : {
      "SIRE_ENTITY_CLASS" : "SPC",
      "SIRE_MENTION_CLASS" : "SPC",
      "SIRE_ENTITY_LEVEL" : "NONE",
      "SIRE_ENTITY_SUBTYPE" : "NONE",
      "SIRE_MENTION_ROLE" : "FURNITURE_PATIO",
      "SIRE_MENTION_TYPE" : "NONE"
   },
   "type" : "FURNITURE_PATIO",
   "begin" : 3221,
   "end" : 3234,
   "inCoref" : false
},
```

   Don't forget to rename the parent mention labels.

   For example, find mentions that specify `"SIRE_ENTITY_SUBTYPE" : "OTHER"`, and then change the value from `OTHER` to `NONE`.

   Change the value of the `SIRE_MENTION_ROLE` and `type` for the mention to the new parent entity type label.

   For example, change the `SIRE_MENTION_ROLE` and `type` values for these mentions from `FURNITURE` to `FURNITURE_NONE`, and the `SIRE_ENTITY_SUBTYPE` to `NONE`.

```
{
   "id" : "Sports_herald.com_be99aca94a7cff5abb74476b844a11b6.en-M75",
   "source" : "IMPORT",
   "properties" : {
      "SIRE_MENTION_CLASS" : "SPC",
      "SIRE_ENTITY_LEVEL" : "NONE",
```

```
      "SIRE_ENTITY_SUBTYPE" : "NONE",
      "SIRE_ENTITY_CLASS" : "SPC",
      "SIRE_MENTION_TYPE" : "NONE",
      "SIRE_MENTION_ROLE" : "FURNITURE_NONE"
   },
   "type" : "FURNITURE_NONE",
   "begin" : 2063,
   "end" : 2071,
   "inCoref" : false
},
```

7. Add annotations for relationships that are missing based on the new flattened entity types.
8. Create a Knowledge Studio workspace, and then upload the converted type system.

   Follow the appropriate steps for uploading a type system based on your Knowledge Studio deployment type:

   - IBM Cloud Adding a type system to the workspace
   - IBM Cloud Pak for Data Adding a type system to the workspace
9. Upload the annotated documents to the workspace. Retain the original file structure of the exported data. Ensure that the compressed file has the same root-level directory as the original exported file, for example.

   Follow the appropriate steps for uploading documents based on your Knowledge Studio deployment type:

   - IBM Cloud Adding documents to a workspace
   - IBM Cloud Pak for Data Adding documents to a workspace
10. From Knowledge Studio, click **Train** to retrain the model.

    For more information, see the appropriate topic for your deployment type:

    - IBM Cloud Training the machine learning model
    - IBM Cloud Pak for Data Training the machine learning model
11. Now, you're ready to export the model from Knowledge Studio and import it to Discovery to use the model as a machine learning enrichment.

    For more information, see Using imported ML models to find custom terms.

Guided tours

You can explore the Discovery interface by taking an interactive tour. Click **Guided tours** from the header of the *My Projects* page to select a tour.

The following tours are available:

- Learn the essentials of Watson Discovery
- Extract meaning from text
- Create and apply dictionaries
- Create a Content Mining project
- Learn the keys to a successful mining setup

Release notes

Release notes for Discovery for IBM Cloud

Learn about features and changes that were included for each release and update of the product software.

IBM Cloud

Note: This information applies only to managed instances of IBM Watson® Discovery that are hosted on IBM Cloud or that were provisioned with IBM Cloud Pak for Data as a Service . For information about releases and updates for installed deployments, see Release notes for IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data.

16 November 2023

APIs for get collection details, list documents, and get document details are now supported in Premium plans of IBM-Cloud managed instances

In Premium plans, the APIs are supported for collections that are created after 16 November 2023. If you want to get information about a collection that was created earlier, trigger a process that runs the conversion step of ingestion on the documents. For example, you can enable the APIs by making changes in the *Identify fields*, *Manage fields*, *CSV settings*, or *Processing settings* (such as OCR settings) pages, or by applying a Smart Document Understanding model to the older collection.

For more information about the new API, see the API reference documentation.

7 November 2023

Preview data for collections

You can preview a document in a collection. To preview data in the advanced document view, navigate to the *Manage collections* page, and click **Preview data** in the collection tile. Alternatively, you can open a collection that you want to preview, and click **Preview data**.

4 October 2023

The optical character recognition (OCR) feature for Hebrew language text in images is a beta feature in Discovery

When OCR is enabled, text extraction and OCR-identified text extraction have limitations for the Hebrew language. These limitations might include the following:

- Inaccurate word order for plain text extraction
- Extracted content in the text and html formats present the text in different word order
- Punctuation and newlines are placed incorrectly in the text
- Text order within a word is reversed depending on the collection settings
- Missing text, text ordered incorrectly, or both might occur when a page contains plain text and image text.

Export labeled data for an entity extractor

You can export the labeled data for an entity extractor for training or building large language models (LLMs). For more information, see Exporting labeled data for an entity extractor.

Find terms that you want to label as entity examples in a document.

You can now search for terms that you want to label as entity examples in a document. You can also find labeled and unlabeled entity examples, and correct any labeling inconsistencies. For more information, see Searching for examples by using keywords .

External enrichment feature to annotate documents with a model of your choice.

Through a webhook interface, you can use custom models or advanced foundation models, and other third-party models for enriching your documents in a collection. For more information, see External enrichment API.

The *Part of Speech* enrichment is no longer available for any project types other than Content Mining

The *Part of Speech* enrichment had been used for dictionary suggestion. However, dictionary suggestion has been updated and it can now work without the *Part of Speech* enrichment applied. For Content Mining projects, the *Part of Speech* enrichment is available as before.

21 September 2023

Updated the tokenizer for all languages

The updated tokenizer might affect the ranking order of results for certain queries. If you observe any ranking differences in your query results, you can reindex the documents in the collection. Discovery tokenizes words both when it ingests and stores data in the index, and at run time when it analyzes queries that are submitted by users. By reindexing the collection, you ensure that your documents are indexed with the same tokenizer that is used for matching queries.

To reindex documents, open the *Manage collection* page, choose a collection, and navigate to the **Enrichments** tab. Select a field to enrich, and then clear the field. Next, click **Apply changes and reprocess** and wait for the documents in the collection to be reprocessed.

15 August 2023

Option to apply or remove a crawl schedule

This option is helpful for easily applying or removing a crawl schedule, and also for stopping a crawl. For more information, see Crawl schedule options.

9 August 2023

You can now specify fields from which to extract content when querying data from the UI

The ability to specify fields allows you to improve the search results when content is not indexed in the default fields. Content might not be indexed in the default fields when you ingest structured files or when you apply a Smart Document Understanding model. For more information, see Excerpt unavailable.

Enrichments in the advanced document view for PDFs are highlighted in distinct colors

When you select multiple enrichments in the advanced document view for PDFs, each enrichment type is highlighted in the document with distinct colors. Overlapping enrichments are also highlighted in a distinct color.

26 July 2023

You can now specify a custom date and time for the crawl schedule

This option is helpful if you want to avoid heavy load on a target system during business hours. For more information, see Crawl schedule options.

10 June 2023

All Entities enrichments use the Entities v2 type system

Natural Language Understanding Entities v1 is no longer supported. IBM Cloud instances that were created before 2 June 2021 and Discovery for IBM Cloud Pak for Data 2.x deployments used version 1 of the Natural Language Understanding Entities type system for English and Korean collections. Now, all collections use only version 2 of the Natural Language Understanding Entities type system.

Classifiers are identified more clearly

The *Enrichments* page lists classifier enrichments as either *text classifier* or *document classifier* enrichments.

16 May 2023

Improved tool for creating Smart Document Understanding (SDU) user-trained models

The SDU tool that you use to annotate documents when you create a user-trained SDU model now uses the React UI framework. This update does not change the behavior of the tool, but does make it more responsive.

You can now define JSON normalizations by using the Collections API

The *Create a collection* and *Update a collection* methods now support the addition of `conversions` and `normalizations` objects that you can specify to apply normalization operations to the documents in the collection. For example, you can define an operation to copy or merge one field to another in the JSON representation of the documents. The `conversions` object defines normalization operations that occur during ingestion and the `normalizations` object defines normalization operations that occur after enrichments are applied. For more information, see the Collections API reference.

31 March 2023

Update to API version

The current API version (v2) is now 2023-03-31. One change was made with this version.

Changed how fields named `document_id` are handled

If you add a JSON file that contains a field named `document_id` to a collection, the field is ignored. The system assigns a new unique document ID to the document when it is added to the index. To assign a document ID to a document regardless of its file type, use the *Update document* method from the API.

Previously, when you uploaded a JSON file with a field named `document_id` from the product user interface or by using the *Add document* API method, the document ID from the file was shown as the `document_id` value in query results. However, a different document ID was assigned to the document, and the assigned ID had to be used for certain other tasks, such as deleting the document. If your application relies on the previous behavior, specify a version number earlier than 2023-03-31, such as `2020-08-30`, in your API calls.

2 March 2023

Now you can specify the types of files to add to a collection

When you connect to an external data source, you can limit the types of files to add to the collection from the external data source. For example, you can choose to add only PDF files from a Box data source.

21 Febuary 2023

Optical character recognition v2 technology is used

The latest version (OCR v2) is used automatically when you enable OCR for English, German, French, Spanish, Dutch, Brazilian Portuguese, and Hebrew collections in all IBM Cloud service plans.

The new optical character recognition model was developed by IBM Research to be better at extracting text from scanned documents and other images that have the following limitations:

- Low quality images due to incorrect scanner settings, insufficient resolution, bad lighting (such as with mobile capture), loss of focus, unaligned pages, and badly printed documents
- Documents with irregular fonts or a variety of colors, font sizes, and backgrounds

The entity extractor limits changed

The number of documents that are allowed in the training data for the Plus plan increased from 100 to 200.

The number of entity types that you can create per plan decreased.

- For Premium plans, the limit changed from 75 to 18.

- For Enterprise plans, the limit changed from 50 to 18.
- For Plus plans, the limit changed from 20 to 12.

The string variation operator now works with phrases

When you include the string variation operator with query input that contains a phrase, the variation is applied to each word in the phrase. For example, `"tom cat"~1` matches `top hat` in addition to `tom cat`. For more information about Discovery Query Language operators, see Query operators.

10 Febuary 2023

Entity extractor is generally available

The *Extract entities* enrichment brings the powerful ability to build a custom type system into Discovery. Use the tool to label entity examples within your industry data to build a machine learning model that Discovery can use to recognize meaningful terms for your business. Already built an entity type system in Knowledge Studio? You can use the corpus from Knowledge Studio as a starting point for your Discovery entity extractor training data. For more information, see Entity extractor.

If you created an entity extractor enrichment for testing purposes when the feature was in beta release, now that it is generally available, it will count toward your custom model limit. The entity extractor enrichment incurs charges whether or not it is applied to a collection.

7 Febuary 2023

Support for hourly crawls was removed

You can no longer choose to crawl a data source every hour. If an existing collection is configured to crawl hourly, you will be prompted to change the scheduled crawl the next time you edit the connector settings.

You can no longer enable FAQ extraction for a collection

The checkbox to enable or disable the beta FAQ extraction feature was removed. FAQ extraction was a beta feature that captured question-and-answer pairs from the data source as it was crawled. FAQ extraction generated a new subdocument for each pair and stored the question in the `title` field and the answer in the `text` field.

You cannot apply FAQ extraction to new collections.

Any existing collections with FAQ extraction enabled retain FAQ documents in their indexes until the collection is reprocessed. At that time, most of the question-and-answer pair subdocuments are deleted. However, any FAQ subdocuments that were generated from HTML or TXT source files remain. If you want to remove these subdocuments, go to the *Manage data* page to delete them. Subdocuments that are generated from one parent document all have the same `metadata.parent_document_id` value.

If you need a way to extract question-and-answer pairs from source documents that use a consistent style and formatting for questions and answers, you can use the Smart Document Understanding tool to annotate the pairs instead. For more information, see Using Smart Document Understanding.

25 January 2023

Set up a Microsoft SharePoint Online data store connector that has *Read* permission

When you create a Microsoft SharePoint Online connector to crawl a SharePoint data source by using Open Authentication v2, the enterprise application that is created by Discovery to make the connection requires *Read* permission only. The enterprise application that was configured for you previously required *Write* permission.

If you want to update an existing connector so that you can use the new Read permission configuration, you must delete your existing enterprise application first.

For more information, see Microsoft SharePoint Online connector.

FAQ extraction deprecation announcement

The beta FAQ extraction feature that detects and extracts question-and-answer pairs from documents is being removed. Support for the feature will end in 1Q 2023.

6 December 2022

Now you can stop a data source crawl

You can stop a crawl that is in progress or that is scheduled to occur in the future. For more information, see Stopping a crawl.

The following item is a known issue:

Box data source scheduled crawls are not updating documents

Due to a problem in the Box Events API, changes that occur between crawls in documents that are stored in Box are not detected and picked up by the Discovery collection during scheduled recrawls. To ensure that your collection is up-to-date, stop and restart the crawl.

1 December 2022

Plus plan supports fewer entity extractors

The maximum number of entity extractors that you can create with a Plus plan decreased from 6 to 3.

12 November 2022

Discovery users might experience issues with documents in collections where OCR is enabled that were added or processed between Nov 1 and Nov 11

Between 1 November and 11 November 2022, some projects with optical character recognition (OCR) enabled, including Document Retrieval for Contracts projects, experienced problems. The problems were related to a new version of the optical character recognition (OCR v2) feature that was enabled automatically for English, German, French, Spanish, Dutch, Brazilian Portuguese, and Hebrew collections during that timeframe. The new version changes sentence boundaries in ways that can negatively impact other functions, including element identification in contracts and the document labeling view in the entity extractor tool.

If you experience any of these issues with documents that were added or processed during this period, revert the version of OCR that is applied to the documents. Starting on 12 November 2022, OCR v1 is applied to all collections where OCR is enabled. To go back to using OCR v1, make a change that will reprocess the affected documents. For example, you can re-add documents that were added during the timeframe to reprocess them. Or you can reprocess an entire collection.

To reprocess a collection, from the *Manage collections* page, open the collection, and then go to the *Processing settings* tab. Expand the *More processing settings* section, set the OCR switch to **Off**, and then set it back to **On**. Click **Apply changes and reprocess** to reprocess your collection.

2 November 2022

A new and improved optical character recognition technology is available

A new version of optical character recognition technology is now available. This latest version (OCR v2) is used automatically when you enable OCR for English, German, French, Spanish, Dutch, Brazilian Portuguese, and Hebrew collections in all IBM Cloud service plans. The new optical character recognition model was developed by IBM Research to be better at extracting text from scanned documents and other images that have the following limitations:

- Low quality images due to incorrect scanner settings, insufficient resolution, bad lighting (such as with mobile capture), loss of focus, unaligned pages, and badly printed documents
- Documents with irregular fonts or a variety of colors, font sizes, and backgrounds

1 November 2022

Entity extractor loads the first 40,000 characters from training data documents

Even extra long documents from the collection that you use to define custom entity examples are loaded into the document view of the tool. However, only the first 40,000 characters, which is approximately 15-20 pages, are displayed. The rest of the file content is truncated. You'll know if your document is truncated because a notification is displayed in the document view. For more information, see Entity extractor.

You can set the passages per document setting to be higher than one

A bug was fixed that prevented you from using the search bar settings in the product user interface to increase the maximum number of passages to return per document. For more information, see How passages are derived.

Improved query aggregation documentation

The documentation that describes the aggregation types that you can specify in the query aggregation parameter was updated. For more information, see Query aggregations.

30 September 2022

Lite plans are no longer available from the London data center

Lite plans are discontinued. You cannot create **new** service instances that use the Lite plan type in any location, including London. Use the new Plus plan and its associated 30-day free trial to explore new features and a simpler way to build that is available with the latest version of the product.

22 September 2022

Plus plan supports more entity extractors

The maximum number of entity extractors that you can create with a Plus plan increased from 3 to 6.

You cannot apply a Smart Document Understanding model to Microsoft Excel files

The quality of structural analysis that can be produced for Excel files is not sufficient. Starting on 22 September 2022, you cannot apply an SDU model to Excel files. This change does not impact Excel files in collections where an SDU model was applied before 22 September 2022.

16 September 2022

In-context document preview is now available for PDF files that are crawled

When you click to view a passage from a search result that is extracted from a PDF document, a document preview page is displayed that shows the returned passage in the context of the original PDF page. The in-context view is available for PDF files to which a Smart Document Understanding model is applied.

15 August 2022

SDKs were updated to reflect the latest API changes.

The following Discovery v2 API changes are now reflected in the SDKs:

- Use the new document classifier API to get, add, update, or delete a document classifier.
- A new document status API is available. You can use it to get a list of the documents in a collection and to get details about a single document.
- You can now get, add, and remove a stop words or expansion list for a collection.
- A `smart_document_understanding` field is returned with the *Get collection* method. This new field specifies whether an SDU model is enabled for the collection and indicates the model type.

- A `similar` parameter is available from the *Query* method. Use it to find documents that are similar to documents of interest to you.
- The `suggested_refinements` parameter of the *Query* method is deprecated. The `suggested_refinements` parameter was used to identify dynamic facets from Premium plan data.

8 August 2022

Larger documents can be crawled

The maximum file sizes that are allowed for crawled documents increased for Premium plans. It also increased for the Box, IBM Cloud Object Storage, and Salesforce connectors. For more information, see File size limits.

2 August 2022

IAM authentication support was added to the IBM Cloud Object Storage connector

You can now choose to authenticate with the IBM Cloud Identity and Access Management (IAM) service. For more information, see IBM Cloud Object Storage.

28 July 2022

API updates

The following changes were made to the Discovery v2 API.

New fields are available:

- A `smart_document_understanding` field is returned with the *Get collection* method. This new field specifies whether an SDU model is enabled for the collection and indicates the model type.
- A `similar` parameter is available from the *Query* method. Use it to find documents that are similar to documents of interest to you.

The `suggested_refinements` parameter of the *Query* method is deprecated. The `suggested_refinements` parameter was used to identify dynamic facets from Premium plan data.

Discovery v1 deprecation announcement

⊖ **Deprecated:** Watson Discovery v1 is being deprecated. Existing clients who use Watson Discovery v1 are asked to migrate to Watson Discovery v2 before the end-of-support date of **11 July 2023**. End of Support means that no v1 instance will work on or after 11 July 2023. For more information about migration, see Getting the most from Discovery.

11 July 2022

The advanced document view highlights even more enrichments

In addition to the built-in *Entities* and *Keywords* enrichments that are recognized by Watson Natural Language Processing models, the advanced document view now highlights the following types of enrichments:

- Custom dictionary terms
- Terms or numbers that match regular expression patterns that you define
- Custom entities and relationships that are defined by Watson Knowledge Studio machine learning and rules-based models
- Custom entities that are defined by using the entity extractor tool that is available as a beta feature

For more information about enrichments that you can add to your documents, see Adding domain-specific resources.

30 June 2022

Watson SDK support change

Support for the following SDKs is provided by the Watson community of developers instead of IBM:

- Go
- Ruby
- Swift
- Unity

For more information, see Watson SDKs.

1 June 2022

The entity extractor tool is now easier to use

The user interface was redesigned to better support the workflow of adding entity types and labeling examples of them. As part of the new design, the bulk labeling feature now is enabled by default, the documents view is easier to find and use, the suggestions pane is more responsive, and you can track metrics scores across multiple training runs. For more information about the entity extractor, see Customizing the terms that Discovery can recognize.

The entity extractor is now available in more plans and languages

The entity extractor beta feature is now available to users of Plus and Enterprise plans in addition to Premium plans. The extractor enrichment is supported for collections in languages other than English.

When you remove a starting URL from a Web crawl connector its associated documents are deleted

The Web crawl connector was updated. Starting with collections that you create after April 2022, if you remove a starting URL from the Web crawl configuration, any indexed documents that were derived from the content of the web page at that URL are deleted with the next crawl. For more information, see Web crawl.

16 May 2022

Added API methods for working with stop words and expansion lists

You can now get, add, and remove a stop words or expansion list for a collection programmatically. For more information, see the Query modifications methods.

13 May 2022

An improved JSON view is available

You can now use keyboard keys to tab through elements in the view. The new JSON view also numbers the occurrences of elements in each JSON object, which makes it easier to keep track of information and to read totals at a glance.

20 April 2022

Analyze API is supported in Enterprise plan deployments

Use the Analyze API to process a JSON file according to a collection's configuration settings, and then return the file for realtime use without storing it in the collection. The Analyze API was supported only in installed deployments previously. For more information, see Analyze API.

A new document status API is available

Use the new document status API to programmatically get a list of the documents in a collection and to get details about a single document. The following notes apply to this release:

- The API is supported for collections that are created after 23 March 2022.

  If you want to get status information about a collection that was created earlier, trigger a process that runs the conversion step of ingestion on the documents. For example, you can enable the API by making changes in the *Identify fields*, *Manage fields*, *CSV settings*, or *Processing settings* (such as OCR or FAQ extraction settings) pages, or by applying a Smart Document Understanding model to the older collection.

- The API is available only from Plus and Enterprise plan instances.

For more information about the new API, see the API reference documentation.

More messages are shown to keep you informed about the status of document processing

An issue was fixed which previously prevented informative messages from being displayed about the status of document conversion and indexing during the ingestion process. Now that the issue is fixed, you might see more messages than usual when you add or reprocess documents. This increase is expected. Nothing you did caused the increase in messages.

6 April 2022

Project tile has a more intuitive menu

The project tile was updated to include an overflow menu that you can use to perform actions such as deleting or renaming a project.

30 March 2022

A new document classifier API is available

Use the new document classifier to programmatically get, add, update, or delete a document classifier. Document classifier methods are supported on installed intances (IBM Cloud Pak for Data) or IBM Cloud-managed Premium or Enterprise plan instances.

For more information about the new API, see the API reference documentation. For more information about adding a document classifier by using the product user interface, see Classifying documents.

21 March 2022

Visualize enrichments found in your documents

When you click to view the passage from a search result, a document preview page is displayed that shows a representation of the original document where the search result was found. For most document types, you can open a new *advanced view* of the document to see useful summary information, such as the number of occurrences of any enrichments that are detected in the document. You also can select one of the enrichments to highlight every occurrence of the element within the document text.

Note: Currently, only the *Entities* and *Keywords* enrichments are listed.

Improved format of search results from PDF documents

When you click to view a passage from a search result that is extracted from a PDF document, a document preview page is displayed that shows the returned passage in the context of the original PDF page.

Note: The in-context view is available for PDF files to which a Smart Document Understanding model is applied. The rich preview does not work on images, meaning it doesn't work on scanned PDF documents. The in-context view is available for PDFs in all languages; however, the enrichment highlighting might be misaligned in some languages.

Tell us what you think

Share your opinions and ideas with us at any time by clicking the **Share feedback** button from the page header of the product user interface.

10 March 2022

Manage the data in a collection from the new *Manage data* page

You can now access the *Manage data* page for a collection from the *Manage collections* navigation pane. Go there to see a list of the documents in your collection and get a quick view of information about the documents. You can also delete documents from a collection with just a few clicks. For more information, see Excluding content from query results.

15 February 2022

An alternative authentication mechanism is available for Microsoft Sharepoint Online connectors

You can now use Open Authentication to sign in to Microsoft SharePoint directly when you configure a new IBM Cloud connector. The *Sign in with Microsoft* option that uses Open Authentication to authenticate with the external data source is a beta feature. For more information, see Microsoft SharePoint Online.

7 January 2022

Upgrade from Plus to Enterprise without help

You can perform an in-place upgrade from a Plus plan to an Enterprise plan. For more information, see Upgrading.

6 December 2021

Crawling web pages with dynamic content is now generally available

The *Execute JavaScript during crawl* feature was introduced as a beta feature, but is now generally available. For more information, see Web crawl.

Capturing the SharePoint ACL information from crawled documents

You can now configure the data source crawl to store ACL information as metadata in the documents that are added to your SharePoint Online collection. For more information, see Microsoft SharePoint Online.

You can add more documents to the training data of the beta entity extractor model

If you added and labeled 20 documents to train a model, and now want to continue to improve the model's performance, you can add more documents. Add the additional documents to the collection that you are using to train the model. After you label the first 20 documents, and the model is up to date with any changes, you can choose to continue labeling documents. The new documents that you added to the collection are loaded. You can label them to augment the training data, and then retrain your model. For more information, see Customizing the terms that Discovery can recognize.

Log out of Discovery

You can log out of the Discovery service instance at any time by clicking **Log out** from the user profile menu that is available from the page header of the product user interface.

18 November 2021

Enterprise plan is now available everywhere

The Enterprise plan is available from all data center locations. Scale and secure your Discovery application with enterprise-grade support and performance, and address more use cases including contract analysis and content mining to explore insights across documents. For more information, see Discovery pricing plans.

11 November 2021

New locations for Enterprise plan now available

The Enterprise plan is available from the Frankfurt, London, Sydney, and Tokyo locations in addition to the Dallas location.

3 November 2021

New Enterprise plan

Scale and secure your Discovery application with enterprise-grade support and performance and address more use cases, including contract analysis and content mining to explore insights across documents. Currently, the Enterprise plan is available only from the Dallas location. For more information, see Discovery pricing plans.

New beta entity extractor enrichment

The *Extract entities* enrichment brings the powerful ability to build a custom type system into Discovery. Use the tool to label entity examples within your industry data to build a machine learning model that Discovery can use to recognize meaningful terms for your business. Currently, this beta feature is available for English-language projects that are created in Premium plan service instances only. For more information, see Customizing the terms that Discovery can recognize.

New *Helpful links* tab

The home page includes a *Helpful links* tab that has quick links to documentation, a community site, and other resources.

Improved field selection choices

When you apply an enrichment to a field or choose a field to use as the source for a facet, the fields that are displayed for you to choose from now include only fields that are valid choices. Previously, the list included fields that were not valid choices.

14 October 2021

New Discovery home page

A new home page is displayed when you start Discovery and gives you quick access to a product overview video, and tours. You can collapse the home page welcome banner to see more projects.

New plan usage section

Stay informed about plan usage and check your usage against the limits for your plan type from the *Plan limits and usage* page. From the product page header, click the user icon. The *Usage* section shows a short summary. Click **View all** to see usage information for all of the plan limit categories.

Change to spelling settings in Search

The spelling correction setting changed from being enabled automatically in new projects to being disabled by default. If you want to alert users when they misspell a term in their query, turn on *Spelling suggestions*. For more information, see Customizing the search bar.

Improved **Guided tours** availability

The **Guided tours** button is now available from the product page header, which make them accessible from anywhere. Previously, it was available from the *My Projects* page only.

1 October 2021

Change to Lite and Advanced plans in all locations

Lite and Advanced plans are discontinued. You cannot create **new** service instances that use the Lite or Advanced plan types in the Dallas, Frankfurt, London, Sydney, Tokyo, and Washington DC locations. Any existing Lite and Advanced plans continue to function properly and continue to be supported. You can upgrade from a Lite plan to an Advanced plan. Use the new Plus plan and its associated 30-day free trial to explore new features

and a simpler way to build that is available with the latest version of the product.

24 September 2021

New scoring for NLU enrichments

Relevance and confidence scores are displayed for NLU enrichments that are returned by search. For example, when you open the JSON view of the document preview from a query result, you can see confidence scores for Entities mentions and relevance scores for Keyword mentions.

9 September 2021

New location for Plus plan

The Plus plan is now available from the Sydney location. Use the new Plus plan and its associated 30-day free trial to explore new features and a simpler way to build that is available with the latest version of the product. For more information, see Getting the most from Discovery.

Change to Lite and Advanced plans in most locations

Lite and Advanced plans are discontinued. You cannot create **new** service instances that use the Lite or Advanced plan types in the Dallas, Frankfurt, London, Sydney, Tokyo, or Washington DC locations. Any existing Lite and Advanced plans continue to function properly and continue to be supported. You can upgrade from a Lite plan to an Advanced plan.

26 August 2021

New locations for the Plus plan

The Plus plan is now available from the London and Washington DC locations, in addition to Dallas, Frankfurt, and Tokyo.

Change to Lite and Advanced plans in some locations

You cannot create **new** service instances that use the Lite or Advanced plan types in the Dallas, Frankfurt, London, Tokyo, or Washington DC locations. Any existing Lite and Advanced plans continue to function properly and continue to be supported. You can upgrade from a Lite plan to an Advanced plan.

New answer finding feature

Answer finding is now generally available for managed deployments. Use answer finding when you want to return a concise answer to a question. For more information, see Answer finding.

16 August 2021

New locations for the Plus plan

The Plus plan is now available from the Frankfurt and Tokyo locations, in addition to Dallas.

Change to Lite and Advanced plans in some locations

Lite and Advanced plans are no longer offered. You cannot create **new** service instances that use the Lite or Advanced plan types in the Dallas, Frankfurt, or Tokyo locations. Any existing Lite and Advanced plans continue to function properly and continue to be supported. You can upgrade from a Lite plan to an Advanced plan.

27 July 2021

Improved document size limit

Document size limit is increased. For Premium plan collections, you can now upload files that are up to 50 MB in size instead of 32 MB. For more

information, see Document limits.

23 July 2021

Improved SharePoint Online connector

The Microsoft SharePoint Online data source connector now accepts any valid Azure Active Directory user ID syntax; the format of the user ID doesn't need to match the `<admin_user>@.onmicrosoft.com` syntax. For more information, see Microsoft SharePoint Online.

16 July 2021

New beta dynamic website web crawl

The Web crawler can now crawl dynamic websites that use JavaScript to render content. If you enable this beta feature, the time it takes to crawl the site increases. For more information, see Web crawl.

23 June 2021

New Plus plan

Use the new Plus plan and its associated 30-day free trial to explore new features and a simpler way to build that is available with the latest version of the product. Currently, the Plus plan is available from the Dallas location. For more information, see Getting the most from Discovery.

Change to Lite and Advanced plans

Lite and Advanced plans are no longer offered. You cannot create **new** service instances that use the Lite or Advanced plan types in the Dallas location. Any existing Lite and Advanced plans continue to function properly and continue to be supported. You can upgrade from a Lite plan to an Advanced plan.

Endpoint deprecation reminder

Change to Discovery API endpoint

As part of work done to fully support Identity and Access Management (IAM) authentication, the endpoint that you use to access your Discovery service programmatically is changing. The old endpoint URLs are deprecated and **will be retired on 26 May 2021**. Update your API calls to use the new URLs.

The pattern for the endpoint URL changed from `gateway-{location}.watsonplatform.net/discovery/api/` to `api.{location}.discovery.watson.cloud.ibm.com/`. The domain, location, and offering identifier are different in the new endpoint. For more information, see Updating endpoint URLs from watsonplatform.net.

If your service instance API credentials use the old endpoint, create a new credential and start using it today. After you update your custom applications to use the new credential, you can delete the old one.

19 March 2021

Improved Web crawl connector

You can use the Web crawl collection type to connect to content that is stored on an internal company website. For more information, see Web crawl.

4 March 2021

New drag and drop feature when uploading

Upload collections now support dragging and dropping documents before and during document upload. For more information, see Uploading data.

New list view for collections

You can view a list of collections that are connected to a particular gateway. For more information, see Viewing collections connected to a gateway.

17 December 2020

Improved date and time display on Activity tab

Each collection now displays the **Next sync scheduled for** date and time on the **Activity** tab of the **Manage collections** page.

New beta FAQ extraction

Released the beta feature FAQ extraction. FAQ extraction automatically extracts question-and-answer pairs from FAQ (frequently asked questions) documents and web pages so that your application returns more precise answers. For more information, see FAQ extraction. For a statement explaining beta features, see Beta features.

3 December 2020

New Content Intelligence

You can now apply the **Contracts** enrichment to a **Document Retrieval** project when you create it. The Contracts enrichment can be used to classify contract terms, parties, effective dates and more within your documents. For more information, see Document Retrieval for Contracts.

10 November 2020

New Box connector

Crawl Box systems. For more information, see Box.

New SharePoint 2016 On-Premises connector

Crawl SharePoint 2016 On-Premises systems. For more information, see SharePoint 2016 On-Premises.

The Box connector does not run on Safari

For more information, see Box connector.

Metadata conversion

If the `metadata` property is converted to an array in the index, the document cannot be deleted by using the *Delete labeled data* API method. For more information, see the API reference.

30 October 2020

New language support for Bosnian, Croatian, Hindi, and Serbian

Basic language support now available for Bosnian, Croatian, Hindi, and Serbian. For more information, see Language support.

New beta Patterns enrichment

The beta release of Patterns enrichment uses pattern induction to help you teach Discovery to recognize patterns in your data. Pattern induction generates extraction patterns from the examples you specify. After you specify a small number of examples, Discovery will suggest additional rules that you verify to complete the pattern. You can use pattern induction as an enrichment or to create a facet. For more information, see Patterns and Creating a facet by identifying a pattern. For a statement explaining beta features, see Beta features.

Change to Document Retrieval projects

In new **Document Retrieval** projects, the `suggested refinements` query setting is now set to `false` by default. It was previously set to `true`.

14 September 2020

New pre-trained model for SDU

A new pre-trained model is available in Smart Document Understanding for Document Retrieval projects. This model is ideal if you need to extract data from documents that include a large number of tables. For more information, see Identifying fields.

30 August 2020

Update to API version

The current API version (v2) is now 2020-08-30. The following change was made with this version:

Change to 'options' object

The List enrichments method no longer returns the `options` object per enrichment. Use the Get enrichment method to return the `options` object for a single enrichment.

16 July 2020

New release for Premium instances

This release is available for Premium instances of Discovery on IBM Cloud created after 16 July 2020. For Premium instances created before that date and for all Lite and Advanced plans, see Getting started with Discovery.

Change to **IBM Cloud Premium**

The Premium plan is now generally available.

New Project-based interface

The project-based UI includes configurations optimized for three common use cases: Document Retrieval, Conversational Search, and Content Mining. For more information, see Creating projects.

New Content Mining app

This entirely new capability of Watson Discovery allows you to find insights in your data when you may not even know the question to ask. The powerful correlation tooling will help you unlock value from large unstructured data sets. For details, see Analyzing your data with the Content Mining application.

New tables as answers

Snippets of text aren't helpful if they are found in a table, so Discovery instead returns a formatted table as an answer if your question is best answered by a table. For more information, see Table retrieval.

New dynamic faceted search feature

Underspecified queries are common. Dynamic Faceted Search automatically categorizes your search results into intelligence facets without training by understanding how they are used in the sentences. See Facets in Document retrieval projects.

New reusable components

You no longer need to build a Discovery application from scratch. We now ship out of the box with reusable, open source, React components. As you configure your Discovery application, you are using the real components. From there you simply deploy to get a custom Discovery application. See Building and deploying components.

New Domain Vocabulary feature

You can build a facet for your users without a Dictionary. Use Domain Vocabulary to build a powerful facet with our understanding of how the data is used in as little as 5 minutes. See Facets.

New relevancy training

You can train at a project level. Discovery ranks the best answer regardless of the data source/collection. See Improving result relevance with

training.

New built-in spelling corrector

Discovery has spelling suggestions built in. See Parameters descriptions.

Improved Autocomplete

Discovery includes autocomplete (type-ahead) for searches, as well as a reusable component for providing this feature to your end users.

New support for 12 languages

Language support for Discovery is now available in 12 additional languages. For the complete list, see Language support.

Cloud Object Storage connector limitation

When connecting to an IBM Cloud® Object Storage data source, only the first 75 buckets for a given credential are displayed.

Current API version

The API version (v2) is `2019-11-29`.

Change to features in this release

Deduplication is not available in this release.

Anomaly Detection is not offered.

IBM Watson® Discovery News is no longer included.

Several Watson Natural Language Understanding enrichments are not available at this time (Entity extraction, Relation extraction, Keyword extraction, Category classification, Concept tagging, Semantic Role extraction, Sentiment analysis, Emotion analysis)

The SharePoint 2016 On-Premises and Box data sources are not available at this time.

Release notes for IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data

Learn about features and changes that were included for each release and update of the product software.

IBM Cloud Pak for Data

> **Note:** This information applies only to instances of IBM Watson® Discovery that are installed on IBM Cloud Pak® for Data. For information about releases and updates for managed deployments, see Release notes for Watson Discovery for IBM Cloud.

For the list of Discovery known issues, see Limitations and known issues in Watson Discovery.

Knowledge Studio for IBM Cloud Pak for Data deprecation announcement

After version 4.7, the operator for IBM Knowledge Studio will no longer be supported and will be removed from the IBM Watson Discovery Cartridge for IBM Cloud Pak for Data and from github.com. The service will not be displayed in the Cloud Pak for Data catalog. This change will not impact existing deployments of the operator.

Migrate your solutions to Watson Discovery, which has powerful custom natural language processing capabilities. Any existing Watson Knowledge Studio for Cloud Pak for Data rules-based or machine learning models can be imported to Watson Discovery and applied to your data as custom enrichments. And the recent release of the custom entities extraction feature brings equivalent function to label and train custom entity models into Watson Discovery. For more information about these features, see Choose enrichments.

For more information about migrating your solutions, see Migrating Knowledge Studio solutions.

4.8.0 release, 29 November 2023

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.8.0 is available.

For a list of new features and bug fixes, see What's new and changed in Watson Discovery

Features that are not available in this release

The following feature is generally available from managed IBM Cloud deployments at the time of this release, but not from installed deployments:

- Answer finding

4.7.3 release, 27 September 2023

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.7.3 is available.

For a list of new features and bug fixes, see What's new and changed in Watson Discovery

Features that are not available in this release

The following feature is generally available from managed IBM Cloud deployments at the time of this release, but not from installed deployments:

- Answer finding

4.7.1 release, 26 July 2023

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.7.1 is available.

For a list of new features and bug fixes, see What's new and changed in Watson Discovery

Features that are not available in this release

The following features are generally available from managed IBM Cloud deployments at the time of this release, but not from installed deployments:

- Answer finding

4.7.0 release, 28 June 2023

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.7.0 is available.

For a list of new features and bug fixes, see What's new and changed in Watson Discovery

Features that are not available in this release

The following features are generally available from managed IBM Cloud deployments at the time of this release, but not from installed deployments:

- Answer finding
- Optical Character Recognition v2

4.6.6 release, 18 May 2023

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data was not refreshed as part of 4.6.6. You can use Discovery 4.6.5 with IBM Cloud Pak for Data 4.6.6.

4.6.5 release, 2 May 2023

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.6.5 is available.

For a list of new features and bug fixes, see What's new and changed in Watson Discovery

Manage the data in a collection from the new *Manage data* page

You can now access a *Manage data* page for a collection. From the new page, you can see a list of the documents in your collection and get a quick view of information about the documents. You can also delete documents from a collection with just a few clicks. For more information, see Excluding content from query results.

You have more control over the data that is crawled by the database connector

When you connect to a database as an external data source, you can now specify the column from which to extract data. If you don't specify the column, a column with text or with a single large object is chosen to be crawled. You can also specify the MIME type of the data in the column that you want to crawl.

Features that are not available in this release

The following features are generally available from managed IBM Cloud deployments at the time of this release, but not from installed deployments:

- Answer finding
- Optical Character Recognition v2

4.6.4 release, 29 March 2023

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data was not refreshed as part of 4.6.4. You can use Discovery 4.6.3 with IBM Cloud Pak for Data 4.6.4 on Red Hat OpenShift Container Platform versions 4.10 or 4.12.

4.6.3 release, 23 february 2023

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.6.3 is available.

For a list of new features and bug fixes, see What's new and changed in Watson Discovery

Features that are not available in this release

The following features are generally available from managed IBM Cloud deployments at the time of this release, but not from installed deployments:

- Answer finding
- *Manage data* page

Important: Back up your data before upgrading to version 4.6.3

Before upgrading to version 4.6.3, you must make a backup of your data. Preserve the backup in a safe location. For more information about backing up your data, see Backing up and restoring data in IBM Cloud Pak for Data . That topic also includes information about restoring your data if that becomes necessary.

4.6.2 release, 30 January 2023

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.6.2 is available.

For a list of new features and bug fixes, see What's new and changed in Watson Discovery

Features that are not available in this release

The following features are generally available from managed IBM Cloud deployments at the time of this release, but not from installed deployments:

- Answer finding
- *Manage data* page

4.6.1 release, December 2022

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data was not refreshed as part of 4.6.1. However, the product documentation was updated with fixes and enhancements.

4.6 release, 30 November 2022

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.6 is available.

For a list of new features and bug fixes, see What's new and changed in Watson Discovery

Features that are not available in this release

The following features are generally available from managed IBM Cloud deployments at the time of this release, but not from installed deployments:

- Answer finding

- *Manage data* page

4.5.3 release, 13 October 2022

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.5.3 is available.

There are no new features in this release. For a list of bug fixes, see What's new and changed in Watson Discovery

Features that are not available in this release

The following features are generally available from managed IBM Cloud deployments at the time of this release, but not from installed deployments:

- Answer finding
- *Manage data* page
- Advanced document view for search results
- The **similar** parameter of the Query method
- The **smart_document_understanding** field in the *Get collection* method response

15 August 2022

SDKs were updated to reflect the latest API changes.

The following Discovery v2 API changes are now reflected in the SDKs:

- Use the new document classifier API to get, add, update, or delete a document classifier.
- A new document status API is available. You can use it to get a list of the documents in a collection and to get details about a single document.
- You can now get, add, and remove a stop words or expansion list for a collection.
- The `suggested_refinements` parameter of the *Query* method is deprecated.

4.5.1 release, 3 August 2022

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.5.1 is available.

There are no new features in this release. For a list of bug fixes, see What's new and changed in Watson Discovery

Features that are not available in this release

The following features are generally available from managed IBM Cloud deployments at the time of this release, but not from installed deployments:

- Answer finding
- *Manage data* page
- Advanced document view for search results
- The **similar** parameter of the Query method
- The **smart_document_understanding** field in the *Get collection* method response

4.5 release, 29 June 2022

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.5 is available.

For a list of new features and bug fixes, see What's new and changed in Watson Discovery

Features that are not available in this release

The following features are generally available from managed IBM Cloud deployments at the time of this release, but not from installed deployments:

- Answer finding
- *Manage data* page
- Advanced document view for search results

4.0.9 release, 25 May 2022

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.0.9 is available.

API usage information is now available from the user interface

You can now get information about analyze API usage from the *Data usage>API usage* page in the product user interface. For more information about the analyze API, see Analyze API.

A new document status API is supported in IBM Cloud Pak® for Data instances

Use the new document status API to programmatically get a list of the documents in a collection and to get details about a single document.

- The API is supported for collections that are created after 23 March 2022.

  If you want to get status information about a collection that was created earlier, trigger a process that runs the conversion step of ingestion on the documents. For example, from the *Activity* page for the collection, click **Recrawl**.

- The API is not supported from the SDKs currently.

For more information about the new API, see the API reference documentation.

Security vulnerabilities were addressed

The following security patches were applied:

- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Node.js
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Apache Xerces
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in OpenSSL
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Google Protocol Buffers

4.0.8 release, 27 April 2022

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.0.8 is available.

The `Development` deployment type was changed to `Starter`

When you install Watson Discovery, you can optionally specify the type of deployment by including the `deploymentType` parameter in your custom resource. The `Development` option is now called the `Starter` option.

The `Development` and `Starter` options are functionally the same, and both values are accepted by the service.

Security vulnerabilities were addressed

The following security patches were applied:

- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Google Protocol Buffers
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Node.js
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Java
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in PostgreSQL
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Kotlin
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Apache POI
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data is affected by a remote code execution in Spring Framework (CVE-2022-22965)

IBM Watson® Discovery for IBM Cloud Private (ICP) for Data 2.2.x End Of Support

Effective 30 April 2022, IBM will withdraw support for the following programs:

- IBM Watson Discovery for ICP for Data 2.2.x
- IBM Watson Discovery for ICP for Data Add-on 2.2.x

For more information, see announcement ENUS921-134.PDF.

4.0.7 release, 30 March 2022

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.0.7 is available.

IBM Cloud Block Storage is now supported

When you install Discovery, you can specify IBM Cloud Block Storage Gold tier (ibmc-block-gold) as your storage class. For more information about the storage class, see Storing data on classic IBM Cloud Block Storage .

Security vulnerabilities were addressed

The following security patches were applied:

- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in NumPy
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Spring
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in FasterXML jackson-databind
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in TensorFlow
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in XStream
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Go

Features that are not available in this release

The following features are generally available from managed IBM Cloud deployments at the time of this release, but not from installed deployments:

- Home page updates
- Answer finding
- *Manage data* page
- Advanced document view for search results

30 March 2020

A new document classifier API is available

Use the new document classifier to programmatically get, add, update, or delete a document classifier. The following notes apply to this release:

- The `enrichments` property of the Document Classifier object is documented as being optional. However, the property is required currently.
- The `field` property in the `federated_classification` object is documented as a string. However, it is currently an array.

For more information about the new API, see the API reference documentation. For more information about adding a document classifier by using the product user interface, see Using the Content Mining application.

The document classifier endpoints are not supported in the SDKs currently.

4.0.6 release, 1 March 2022

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.0.6 is available.

Multitenancy is now supported

An administrator can now create up to 10 instances of the Discovery service per deployment, which means that more teams can work on discrete Discovery projects at the same time.

Simpler installation and management of custom connectors

The `manage_custom_crawler.sh` script was improved to make it easier for you to install and manage your custom connectors in a multitenant environment. For more information, see Installing a custom crawler.

Security vulnerabilities were addressed

The following security patches were applied:

- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Java
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Logback
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Apache Log4j

Features that are not available in this release

The following features are generally available from managed IBM Cloud deployments at the time of this release, but not from installed deployments:

- Home page updates
- Answer finding
- Access to guided tours from the page header

4.0.5 release, 26 January 2022

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.0.5 is available.

A security vulnerability was addressed

The following security patch was applied: Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Apache Log4j

Features that are not available in this release

The following features are generally available from IBM Cloud deployments at the time of this release, but not from installed deployments:

- Home page updates
- Answer finding
- Guided tours

4.0.4 release, 20 December 2021

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.0.4 is available.

Guided tours are available

Access guided tours from anywhere in the product user interface by clicking the **Guided tours** button in the page header.

Security vulnerabilities were addressed

The following security patches were applied:

- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in LibTIFF
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in TensorFlow
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Node.js
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Netty
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Apache Log4j
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Apache Log4j 1.2

Features that are not available in this release

The following features are generally available from IBM Cloud deployments at the time of this release, but not from installed deployments:

- Home page updates

- Answer finding

4.0.3 release, 30 November 2021

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.0.3 is available.

Another storage option is supported

IBM Spectrum Scale Container Native storage is now supported in addition to Red Hat OpenShift Container Storage and Portworx.

Microsoft SharePoint Online data source improvement

The *Sharepoint Online* data source now supports crawling your data as a service principal, which means you can access your data without disabling multifactor authentication. For more information, see Microsoft Sharepoint Online.

Microsoft Windows File System improvements

Extra configuration options mean you can specify the following information:

- The types of files (by file extension) to include or exclude from a crawl of a Windows directory.
- The character encoding of the data to be crawled. Typically, the encoding is detected automatically. However, you can choose to specify the character encoding as a Java character set yourself.

For more information, see Windows File System.

Field selection is improved

When you apply an enrichment to a field or choose a field to use as the source for a facet, the fields that are displayed for you to choose from now shows only fields that are valid choices.

Search settings change

The spelling correction setting changed from being enabled automatically in new projects to being disabled by default. If you want to alert users when they misspell a term in their query, turn on *Spelling suggestions*. For more information, see Customizing the search bar.

A Salesforce crawling issue was fixed

Previously, Discovery had an issue where it timed out before it crawled some of the object types in a Salesforce collection. If your collection is configured to crawl the following object types, run a full data source crawl to make sure that your collection contains the most up-to-date data from all of the objects in your Salesforce data source:

- Attachment
- ContentVersion
- Document

Security vulnerabilities were addressed

The following security patches were applied:

- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Node.js
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Axios
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Python Pillow
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Apache Commons Compress
- Security Bulletin: IBM Watson Discovery for IBM Cloud Pak for Data affected by vulnerability in Java

Features that are not available in this release

The following features are available from IBM Cloud deployments at the time of this release, but not from installed deployments:

- Home page updates
- Guided tours
- Answer finding

4.0.2 release, 5 October 2021

IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.0.2 is available.

Support for newer platform software

IBM Cloud Pak® for Data 4.0.2 can be installed on Red Hat® OpenShift® on IBM Cloud® 4.8.

New scoring for NLU enrichments

Relevance and confidence scores are displayed for NLU enrichments that are returned by search. For example, when you open the JSON view of the document preview from a query result, you can see confidence scores for Entities mentions and relevance scores for Keyword mentions.

Improved Web crawl

The *Web crawl* data source supports more customization options, including the ability to ignore a site's robots.txt file. For more information, see Web crawl.

New upgrade support

The 4.0.2 release supports in-place upgrade from IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data 4.0.0. For more information, see Upgrading Watson Discovery to a newer 4.0 refresh

IBM Cloud Private End Of Support

Effective 30 September 2021, IBM withdrew support for the following programs:

- IBM Watson Assistant Discovery Extension for IBM Cloud Private 2.1.0–2.1.4
- IBM Watson Discovery for ICP for Data 2.1.0–2.1.4
- IBM Watson Discovery for ICP for Data Add-on 2.1.0–2.1.4

For more information, see announcements ENUS921-005.PDF and ENUSLP21-0099.PDF.

4 release, 13 July 2021

New version now available

Discovery for Cloud Pak for Data 4 is available

This release is supported on IBM Cloud Pak® for Data 4.0.0.

Change to service name

The new name is IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data.

New Smart Document Understanding (SDU) predefined model

When you identify fields, instead of annotating documents with the SDU tool, you can choose to use a pretrained model. The pretrained model applies a non-customizable model that automatically extracts text and identifies tables, lists, and sections.

Improved contract analysis

To enable the Contracts enrichment that recognizes and tags contract-related concepts in your data, you can choose to create a Document Retrieval project type, and then select **Apply contracts enrichment**. You no longer need to use an installation override YAML file to enable it. This change also means that you can choose which Document Retrieval projects use the Contracts enrichment; it is not applied to all Document Retrieval projects automatically.

New LDAP directory data source

Connect to data that is stored in an external directory that supports the Lightweight Directory Access Protocol (LDAP), such as a corporate email directory. As the directory data is added to your collection, Discovery interprets and stores key attributes of each record, such as department and location information. Later, you can find relevant records by filtering on these attribute categories. For more information, see LDAP directory.

Improved SharePoint OnPrem connection process

The steps you follow to connect to a SharePoint instance that is hosted on-premises were simplified. You no longer need to deploy a web services package on the SharePoint server before you can connect to the SharePoint OnPrem data source. For more information, see SharePoint OnPrem.

New Salesforce proxy support

You can now connect to a Salesforce data source when using a proxy server. For more information, see Salesforce.

Improved custom connector improvements

Support was added for Optical character recognition (OCR)

Support was added for Document-level security

For more information about the custom connector, see Building a Cloud Pak for Data custom connector.

Change to Dynamic Faceted Search

Support for *Dynamic Faceted Search* and its associated `suggested_refinements` API query parameter was removed.

2.2.1 release, 26 February 2021

New release now available

IBM Watson™ Discovery for IBM Cloud Pak for Data version 2.2.1 is available.

Support for upgrade

Discovery for Cloud Pak for Data supports an in-place upgrade from version 2.2.0 to 2.2.1 so that you do not need to manually uninstall an earlier version and then install the latest version of the service. For more information, see Upgrading Discovery for Cloud Pak for Data.

New SDK download support

You can now download the custom connector SDK package from your Discovery for Cloud Pak for Data cluster, instead of retrieving the images and the SDK package from the Docker registry. For more information, see Downloading the custom-crawler-docs.zip file in Discovery 2.2.1 and later.

Change to Invoices and Purchase orders

`Invoices` and `Purchase orders` models can no longer be enabled in the tooling. If you need these models, please contact IBM Cloud Support to obtain instructions for enabling these models.

Change to Contracts enrichment tables

In a `Document Retrieval` project that has the `Contracts` enrichment applied, tables are not included inside the `contracts` field, as they were previously in projects that had the `Contracts` enrichment enabled. Tables will continue to be included in a separate `tables` field when the `Table Understanding` enrichment is applied.

Change to support for Oracle Database 11g and Postgres 9.5

Support for connecting to Oracle Database 11g was removed because the vendor ended version support on 31 December 2020.

Support for connecting to Postgres 9.5 was removed because the vendor ended version support on 11 February 2021.

2.2.0 release, 8 December 2020

New release now available

IBM Watson™ Discovery for IBM Cloud Pak for Data version 2.2 is available.

Discovery for Cloud Pak for Data now works with IIBM Cloud Pak® for Data 3.5.

New support for Notes attachments

Added support for attachments in the Notes data source. For more information, see Notes

New web crawl scheduling option

You can specify the exact time that you would like your crawls to run for any data source, giving you the flexibility to run them at the times you prefer. For more information, see Configuring Cloud Pak for Data data sources.

New Facet creation in Content Miner

You can now create Facet groups in a Content Miner application.

New custom crawler creation

Added the option to create your own custom crawler plug-in. For more information, see Building a Cloud Pak for Data crawler plug-in **Note:** Any custom code used with Watson Discovery is the responsibility of the developer and is not covered by IBM support.

Change to Dynamic Facets

Dynamic Facets are no longer enabled by default in Document Retrieval projects.

2.1.4 release, 2 September 2020

New release now available

IBM Watson® Discovery for IBM Cloud Pak® for Data version 2.1.4 is available.

New Notes connector

Crawl Notes version 9.0.1 systems. For more information, see Notes connector.

New **Enable proxy settings** in multiple connectors

You can now select the option to enable proxy settings in Box, Microsoft SharePoint Online, and Microsoft SharePoint OnPrem connectors.

New options for Database connector

Added support for multiple tables and the Row filter option to the Database connector.

New authentication types for Web crawler

You can select from three new authentication types in Web crawler: Basic authentication, NTLM authentication, and FORM authentication.

New Analyze API usage monitoring

You can now monitor the usage of the Analyze API using the tooling. For more information, see Monitoring usage.

30 August 2020

Update to API version

The current API version (v2) is now 2020-08-30. The following change was made with this version:

Change to 'options' object

The List enrichments method no longer returns the `options` object per enrichment. Use the Get enrichment method to return the `options` object for a single enrichment.

2.1.3 release, 19 June 2020

New release now available

IBM Watson® Discovery for IBM Cloud Pak® for Data version 2.1.3 is available.

Discovery for Cloud Pak for Data now works with IBM Cloud Pak® for Data 3.0.1.

New Finnish and Hebrew language support

Added basic support for Finnish and Hebrew. For more information, see Language support.

Change to Analyze endpoint

The Analyze endpoint, which supports stateless document ingestion workflows. For details, see the Analyze API. The Analyze API supports JSON documents only. Use of the Analyze API affects license usage.

New options for Content Miner

The content mining application includes two new options: Cyclic time scale on the *Time series* dashboard, and the *Contextual view* tab.

New shortcut for Content Mining projects

For **Content Mining** projects only, the **Improve and customize** page includes a shortcut: the **Launch application** button. Previously, you were required to open the **Integrate and deploy** page, select the **Launch application** tab, and click the **Launch** button.

Improved segment limit

The segment limit when splitting documents has been increased to 1,000. For details, see Split documents to make query results more succinct.

Improved Filenet connector

The Filenet connector has document level security.

New beta Curations feature

You can specify up to 1,000 curations. For details about this beta feature, see Curations.

Fixed defects in the 2.1.3 release

In versions 2.1.2, 2.1.1, and 2.1.0, PNG, TIFF, and JPG individual image files are not scanned, and no text is extracted from those files. PNG, TIFF, and JPEG images embedded in PDF, Word, PowerPoint, and Excel files are also not scanned, and no text is extracted from those image files.

2.1.2 release, 31 March 2020

New release now available

IBM Watson® Discovery for IBM Cloud Pak® for Data version 2.1.2 is available.

New **IBM FileNet connector**

You can now crawl IBM FileNet systems. For more information, see FileNet connector.

New Swedish, Norwegian, and Danish language support

Added basic support for Swedish, Norwegian (Bokmål and Nynorsk), and Danish. For more information, see Language support.

Change to Advanced rules models enrichment

The Advanced rules models enrichment is now GA.

New document preview for search results

You can now view your search results in a document preview for the following source documents: PDF, Word, PowerPoint, Excel, and all image files. See supported file types for the list of image files. This view makes it easier for you to see search results as highlighted passages within the text of the original document, making the context clearer.

New proxy support for Web Crawl

Support was added to the Web Crawl connector for proxy support.

Change to empty aggregations parameter

Running a query with an empty `aggregations` parameter returns zero aggregations in the response.

Change to Postgres support

Support for connecting to Postgres 9.4 was removed because the vendor ended version support was ended by the vendor on 13 February 2020.

Fixed the following defects in the 2.1.2 release

When installing Discovery for Cloud Pak for Data on OpenShift, the `ranker-rest` service might intermittently fail to startup, due to an incompatible jar in the `classpath`.

When you upload documents to a collection with existing documents, a `Documents uploaded!` message displays on the **Activity** page, but no further processing status displays until the number of documents increases.

Running a query with an empty `aggregations` parameter returns an empty aggregations array.

Deprovisioning a IBM Watson® Discovery for IBM Cloud Pak® for Data Instance will not delete the underlying data. Delete the collections and documents manually.

2.1.1 release, 24 January 2020

New release now available

IBM Watson® Discovery for IBM Cloud Pak® for Data version 2.1.1 is available.

Fixed the following defects in the 2.1.1 release:

In Document Retrieval project types, when you perform an empty search, and the search results source is set to `passages,` the query results will display `excerpt unavailable` in the Project workspace.

When visiting the Storybook links on the Integrate and deploy page, the links do not go to the correct location. Please visit Storybook instead to view documentation.

If you are using Smart Document Understanding, two variables no longer need to be set during installation or reinstallation. For more information, see Environment variable settings for Smart Document Understanding.

Discovery for Content Intelligence and Table Understanding enrichments are configured out of the box to be applied on a field named `html`. When a user uploads a JSON document without a root-level field named `html`, these enrichments will not yield results in the index. To run the enrichments on this kind of JSON documents, users must re-configure the enrichments to run on an existing field (or fields) in the JSON document.

2.1.0 release, 27 November 2019

New release now available

IBM Watson® Discovery for IBM Cloud Pak® for Data version 2.1.0 is available.

Discovery for Cloud Pak for Data now works with IBM Cloud Pak® for Data 2.5.0.0.

New Project-based interface

Test your application like an end-user would with the **Document retrieval**, **Conversational Search**, and **Content Mining** project types. For more information, see Creating projects.

New Content Mining app

Build an end user interface for extracting insights proactively from your entire corpus. For more information, see Analyzing your data with the Content Mining application.

New Content Intelligence add-on

Option to enrich your documents with pre-built domain knowledge for Contracts. For more information, see Document Retrieval for Contracts.

New reusable components

Use reusable components to quickly build your application with Discovery. We ship an autocomplete, rich preview, results and facets component. For more information, see Building and deploying components.

New Czech, Polish, Romanian, Russian, and Slovak language support

Basic support for Czech, Slovak, Russian, Polish and Romanian is added. For more information, see Language support.

New built-in table understanding

Extract tables from your documents without training, and optionally return tables as answers to natural language queries. For more information, see Understanding tables.

New SDK connector

Build custom connectors your Discovery users can use to build their own applications. For more information, see Building and implementing a custom connector.

New pre-built sample project

The sample project is preloaded with data, so you can learn about Discovery. For more information, see Getting started with Watson Discovery.

New passage retrieval

Will return the most relevant passages from your documents, plus you can specify the number of passages returned per document. See Passages.

New project-level querying and relevancy training

Query multiple collections at once including relevance training.

Improved Web crawl connector

Additional options now available for the **Web crawl connector** - For more information, see Web crawl.

New Local File System connector

Crawl Linux or other file systems. For more information, see Local file system

New dynamic Facets

Automatically generate facets based on the understanding of your data. For more information, see Facets.

New Dictionary suggestions

Dictionary terms are suggested based on your content. For more information, see Dictionary.

New beta Curations

Specify a particular result for a given query. For more information, see the API reference.

2.0.1 release, 30 August 2019

New release now available

IBM Watson® Discovery for IBM Cloud Pak® for Data version 2.0.1 is available.

Discovery for Cloud Pak for Data now works with IBM Cloud Pak® for Data 2.1.0.1.

New Windows File System and Database connectors

Added the Windows File System and Database connectors. For more information, see Database connector and Windows File System connector.

New Chinese language support

Added support for Traditional Chinese. For more information, see Language support.

New FISMA support

Federal Information Security Management Act (FISMA) support is available for IBM Watson® Discovery for IBM Cloud Pak® for Data offerings purchased on or after August 30, 2019. FISMA support is also available to those who purchased the June 28, 2019 version and upgrade to the August 30, 2019 version. IBM Watson® Discovery for IBM Cloud Pak® for Data is FISMA High Ready.

New Classifier enrichment

Released the Classifier enrichment. For more information, see Classifier.

New Red Hat OpenShift support

Added support for installing IBM Cloud Pak® for Data on Red Hat OpenShift.

Fixed the following defects in Discovery for Cloud Pak for Data offerings purchased on or after August 30, 2019

During an active web crawl, if you add an enrichment, then click the **Recrawl collection** button on the **Activity** page, the collection will stop processing. If the collection does not return to a Syncing state on its own, clicking the **Recrawl collection** button an additional time might be required.

While training a collection in the tooling , if you rate the relevancy of a result (for example, as `Relevant`), then switch to the opposite rating (`Not relevant`), the page may go blank. To restore the page, refresh the browser. Your updated rating will be retained.

Chinese, Japanese, and Korean language Microsoft Word, Excel, and PowerPoint documents will not display correctly in the index or the Smart Document Understanding editor.

If you upload a zip, gzip, or tar file to your collection, and that file contains multiple files/file types supported by Smart Document Understanding (PDF, Word, Excel, PowerPoint, PNG, TIFF, JPEG), only one of the files in that zip, gzip, or tar file will be available for training in the SDU editor (unless the SDU document limit has already been met). All of the documents will be available in the index. Unzip the file before uploading to avoid this issue.

Query expansion and autocomplete return the wrong error code when the `collection_id` is invalid. Query expansion will return a `500` error code instead of a `404`. Autocomplete will return a `400` when the `collection_id` is invalid and the `prefix` parameter isn't set. It should also return a `404`.

When crawling Microsoft SharePoint 2019 collections, only HTML documents will be crawled and indexed. This is a SharePoint issue with how it processes mime-types. See this Microsoft blog post for a workaround.

If you delete an installation of the Discovery for Cloud Pak for Data add-on, the instance will not uninstall completely and your re-installation will fail. See the Discovery for Cloud Pak for Data Readme for post-cleanup steps.

If a JSON document that contains nested JSON objects is ingested, the nested JSON will be indexed as a JSON string.

2.0.0, General Availability (GA) release, 28 June 2019

Discovery for Cloud Pak for Data now available

The IBM Watson® Discovery for IBM Cloud Pak® for Data service brings the cognitive capabilities of IBM Watson® Discovery to the IBM Cloud Pak® for Data platform.

Power your assistant with answers from web resources

In this tutorial, you will use the Watson Discovery and watsonx Assistant services to create a virtual assistant that can answer questions about the latest research from the US Federal Reserve. The assistant will answer questions by using up-to-date, existing research publications from the Federal Reserve Economic Data (FRED) website.

IBM Cloud

Note: Follow this tutorial only if you are using a managed deployment.

Learning objectives

By the time you finish the tutorial, you will understand how to:

- Create an action in watsonx Assistant that can recognize questions about a particular subject.
- Create a Conversational Search project in Discovery.
- Add a web crawl data source to your project.
- Connect your watsonx Assistant action to a search extension that gets answers from your Discovery project.
- Use your assistant to return answers that it retrieves from the website.

Duration

This tutorial will take approximately 2 to 3 hours to complete.

Prerequisite

1. Before you begin, you must set up a paid account with IBM Cloud.

   You can complete this tutorial at no cost by using a Plus plan, which offers a 30-day trial at no cost. However, to create a Plus plan instance of the service, you must have a paid account (where you provide credit card details). For more information about creating a paid account, see Upgrading your account.

2. Create a Plus plan Discovery service instance.

   Go to the Discovery resource page in the IBM Cloud catalog and create a Plus plan service instance.

Important: If you decide to stop using the Plus plan and don't want to pay for it, delete the Plus plan service instance before the 30-day trial period ends.

Step 1: Create an assistant

For this tutorial, you will create an assistant with a single action. First, you must create a watsonx Assistant service instance.

Both Lite and Trial plan watsonx Assistant service instances are available at no cost. You will create a Trial plan because a Plus or higher plan is required to add a search skill to an assistant and the Trial plan includes all Plus plan features. The Lite plan does not.

1. Create a Trial plan watsonx Assistant service instance in the same data location where the Discovery service instance is hosted, such as Dallas.
2. From the watsonx Assistant plan service page in IBM Cloud, click **Launch watsonx Assistant**.

   The watsonx Assistant product user interface is displayed where you can create your first assistant.
3. Add `FRED research` as the assistant name, and then click **Next**.

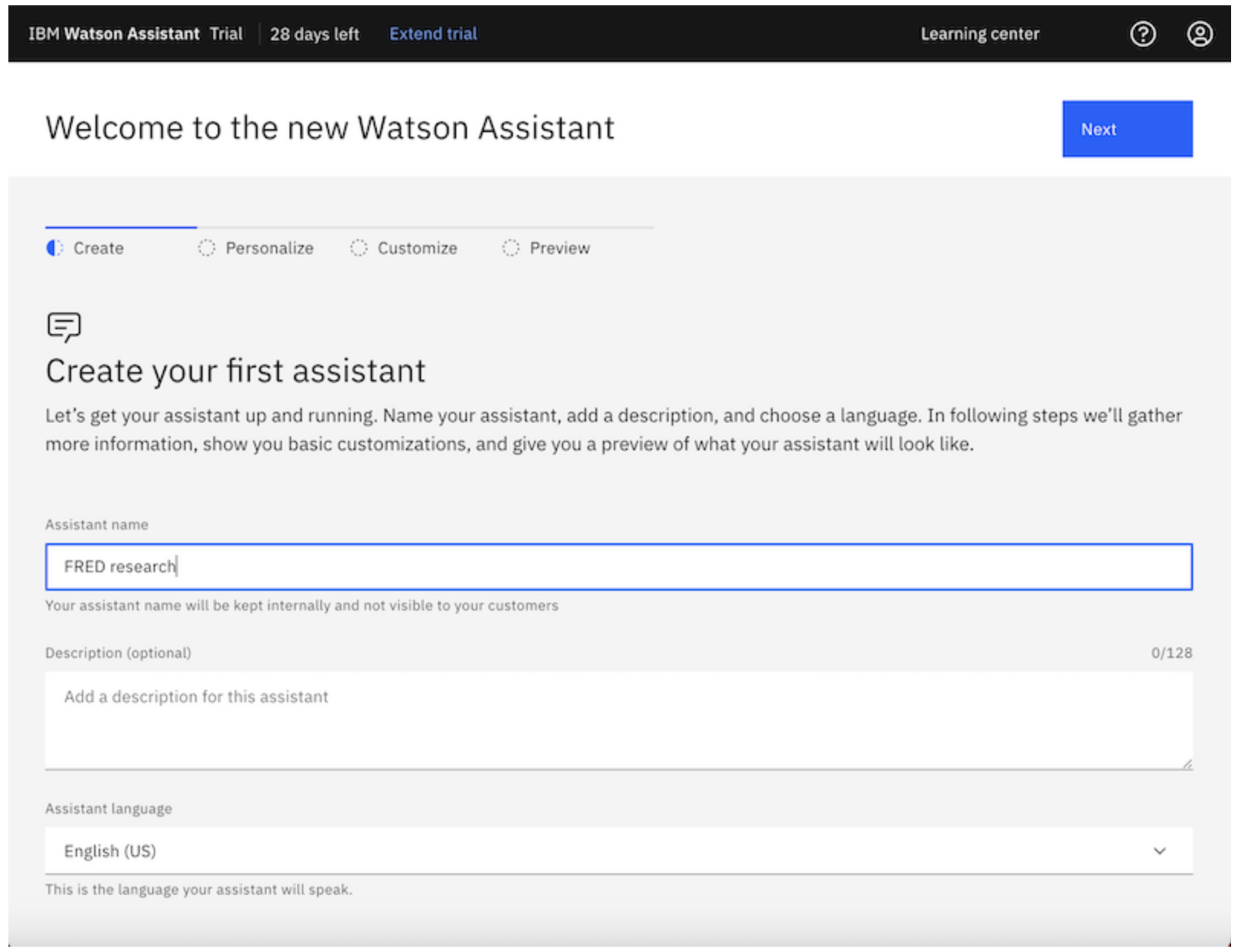

Figure 1. Watson Assistant welcome page

4. Fill out the fields to share information about you and your assistant, and then click **Next**.

   In the *Which statement describes your needs best* field, choose **I'm using Watson Assistant to complete a course or certification.** .

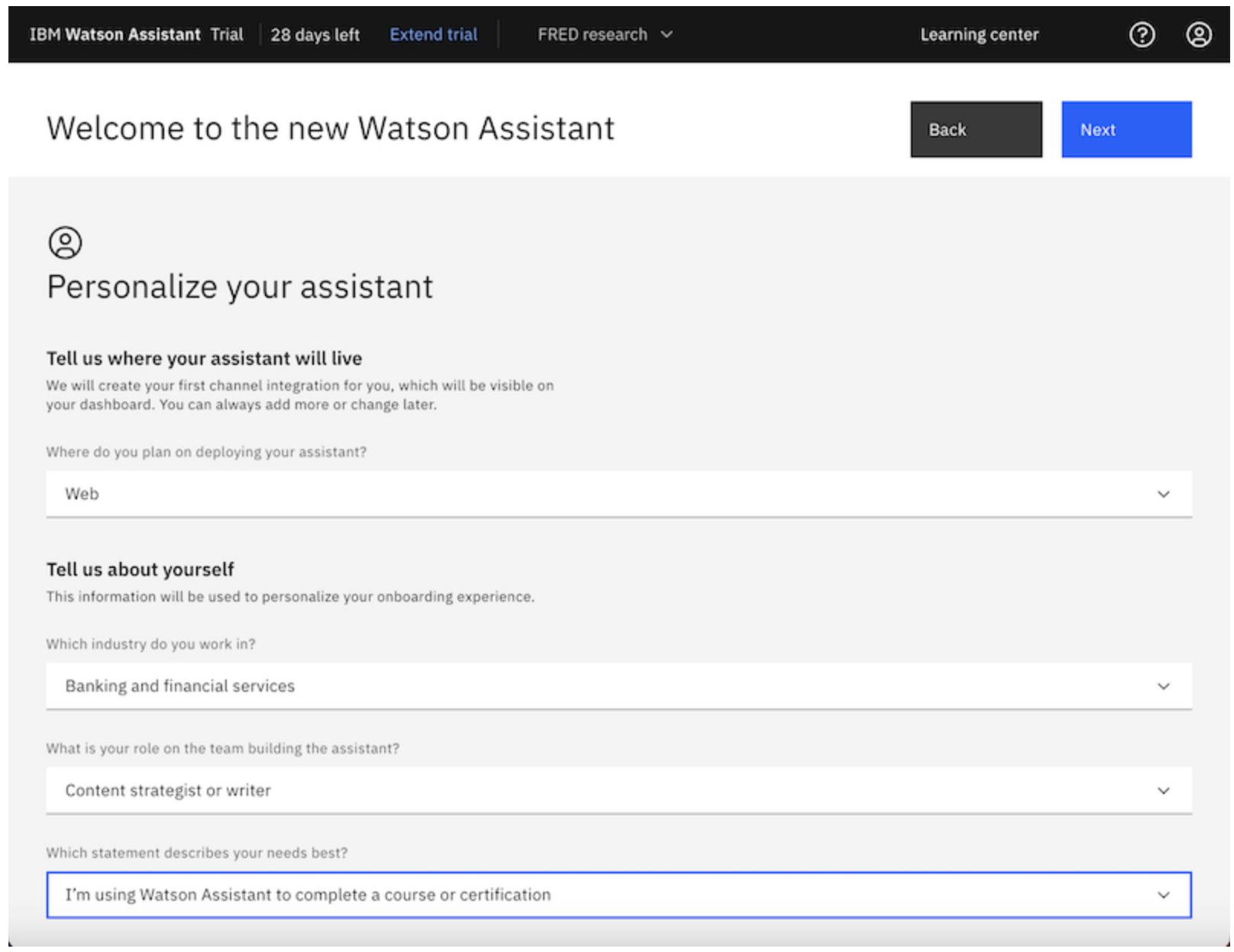

Figure 2. Assistant details page

5. When you create an assistant, a web chat application is created for you automatically.

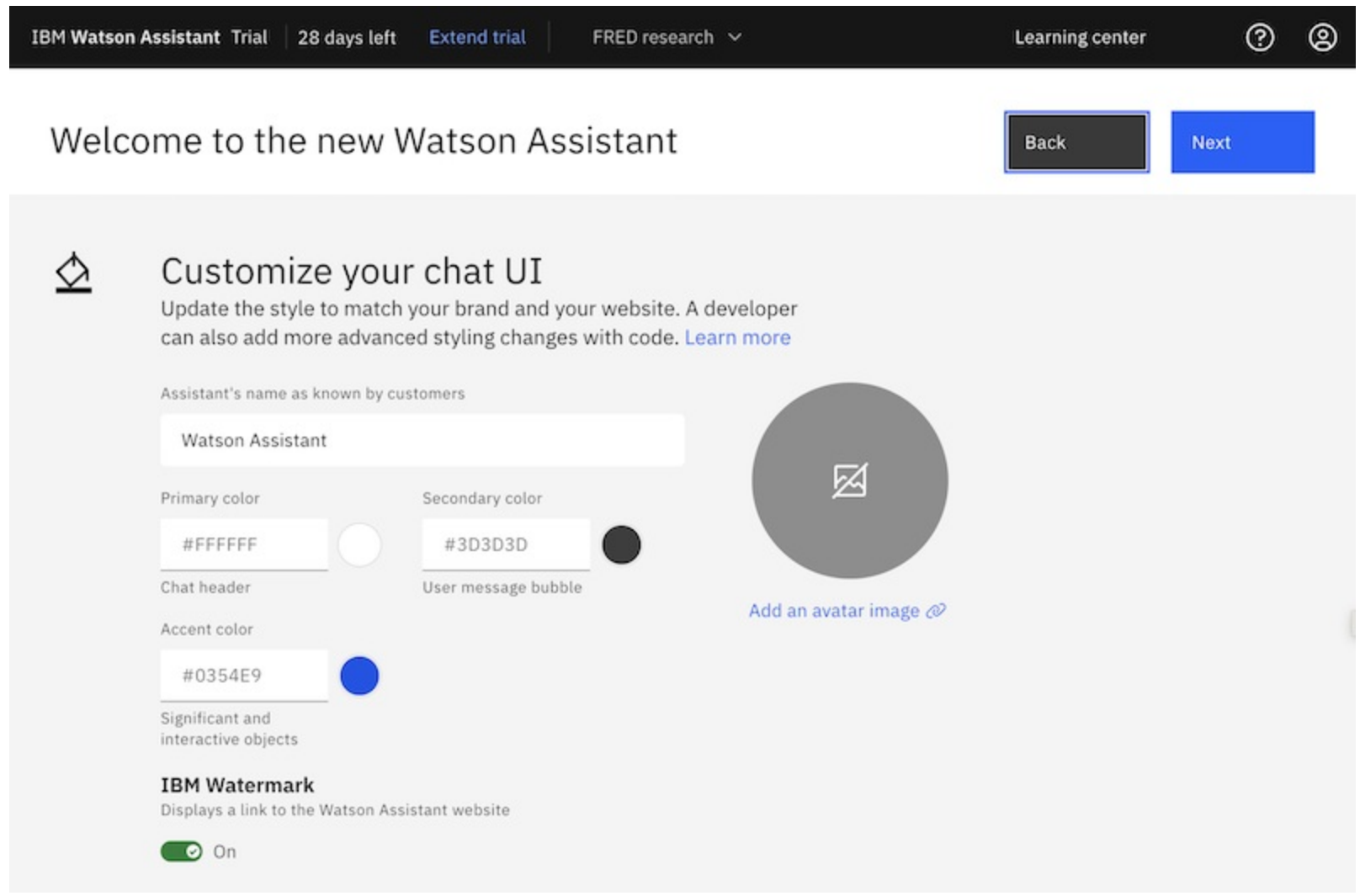

Figure 3. Web chat settings

6. Click **Next** to accept the default style for the web chat.

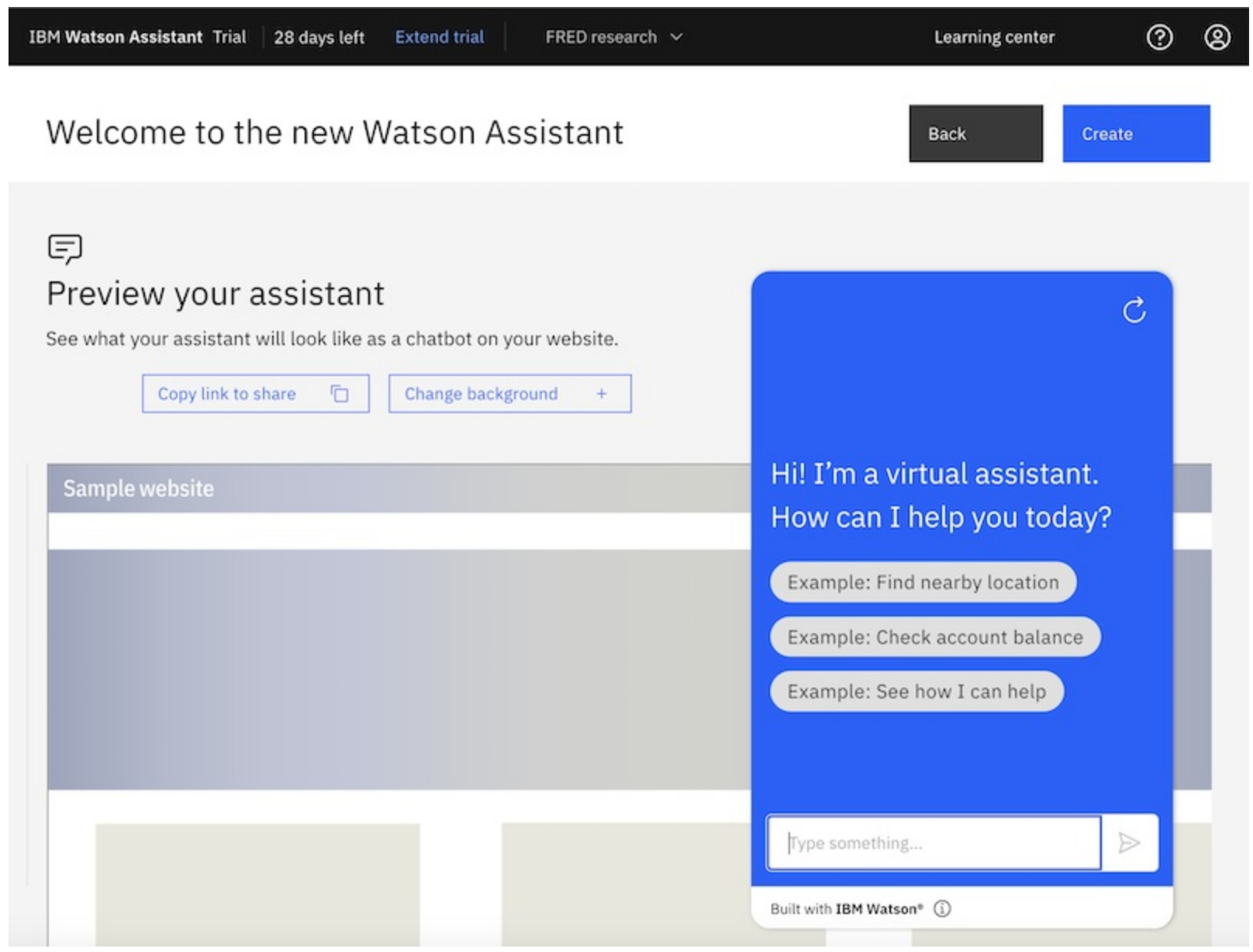

Figure 4. Web chat preview

A preview of the web chat as it would be displayed in a web page is shown.

7. Click **Create** to create the assistant and the corresponding web chat app.

After a congratulatory message, the home page for your new assistant is displayed.

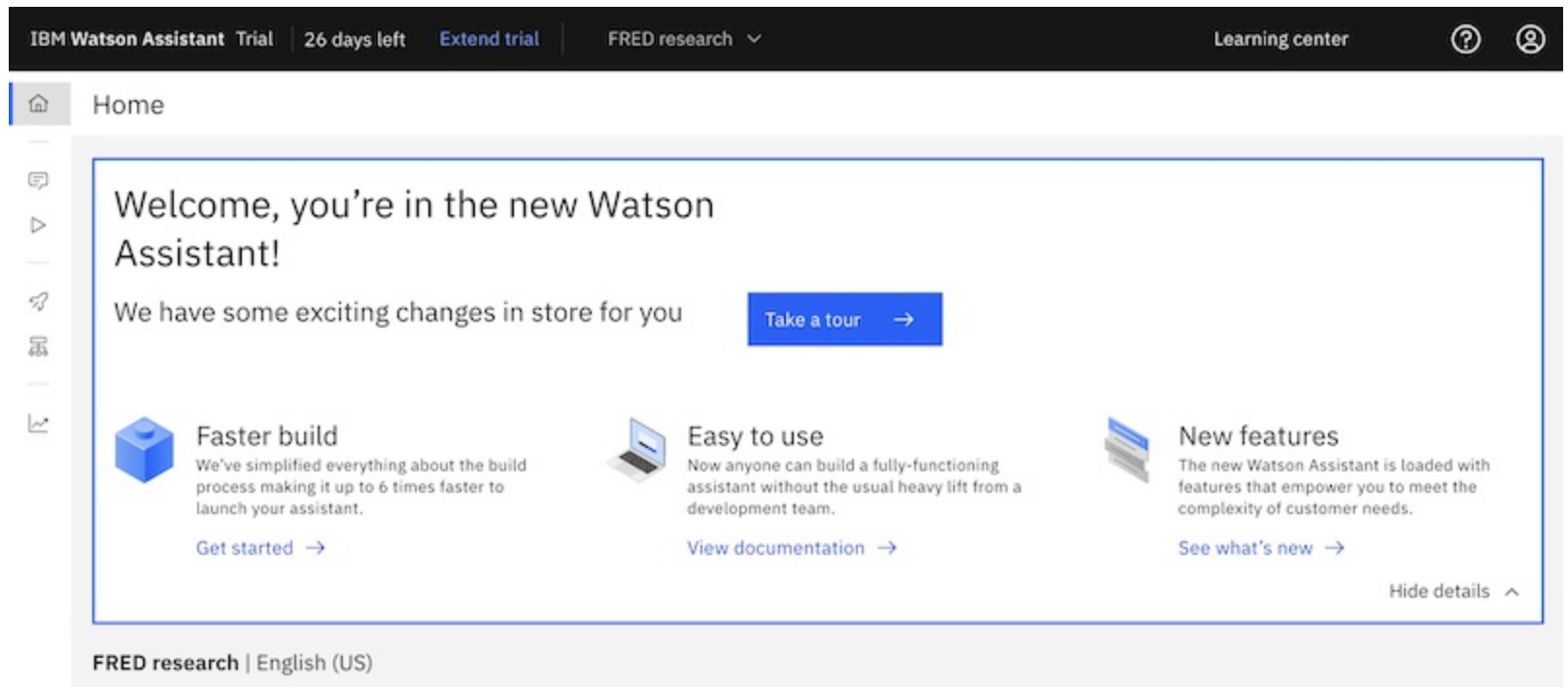

Figure 5. Assistant home page

Step 2: Create an action

Create a single action that can recognize questions about the latest research papers from the US Federal Reserve Economic Data (FRED) website.

In a real world scenario, you might want your assistant to answer questions about the products in your catalog or about insurance plan options or anything else. You can complete similar steps to teach the assistant to recognize when a customer is asking about a particular subject.

1. From the navigation panel, click **Actions**.

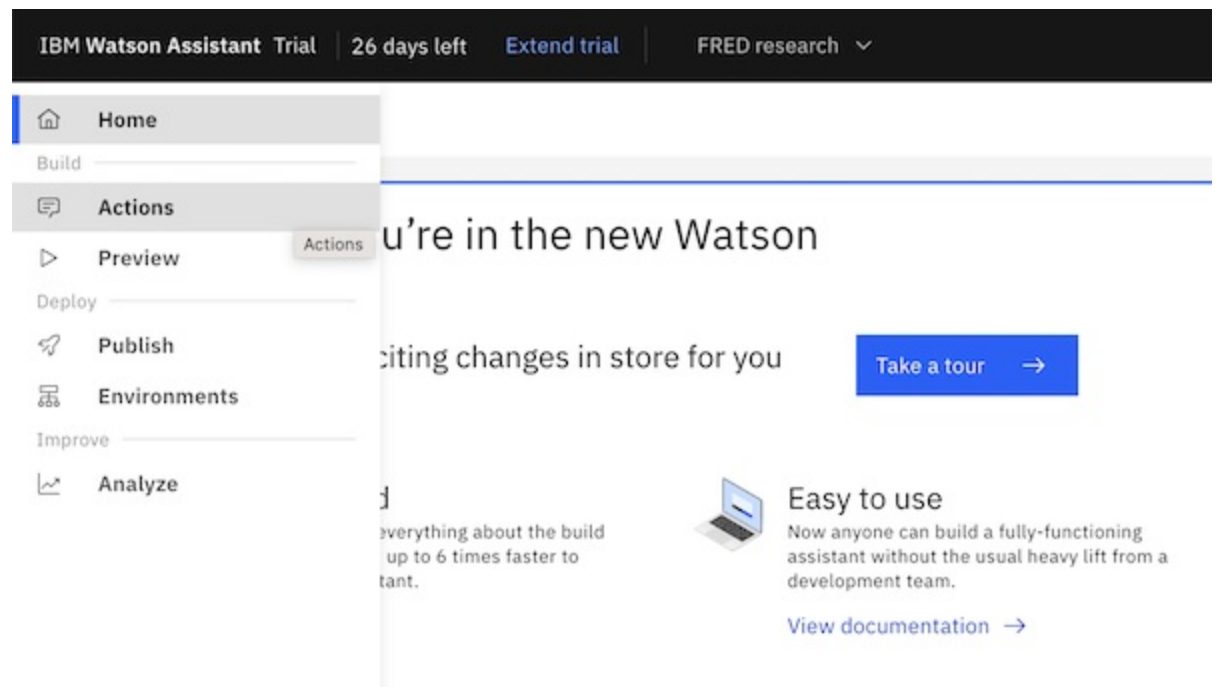

Figure 6. Actions menu

The Actions page is displayed.

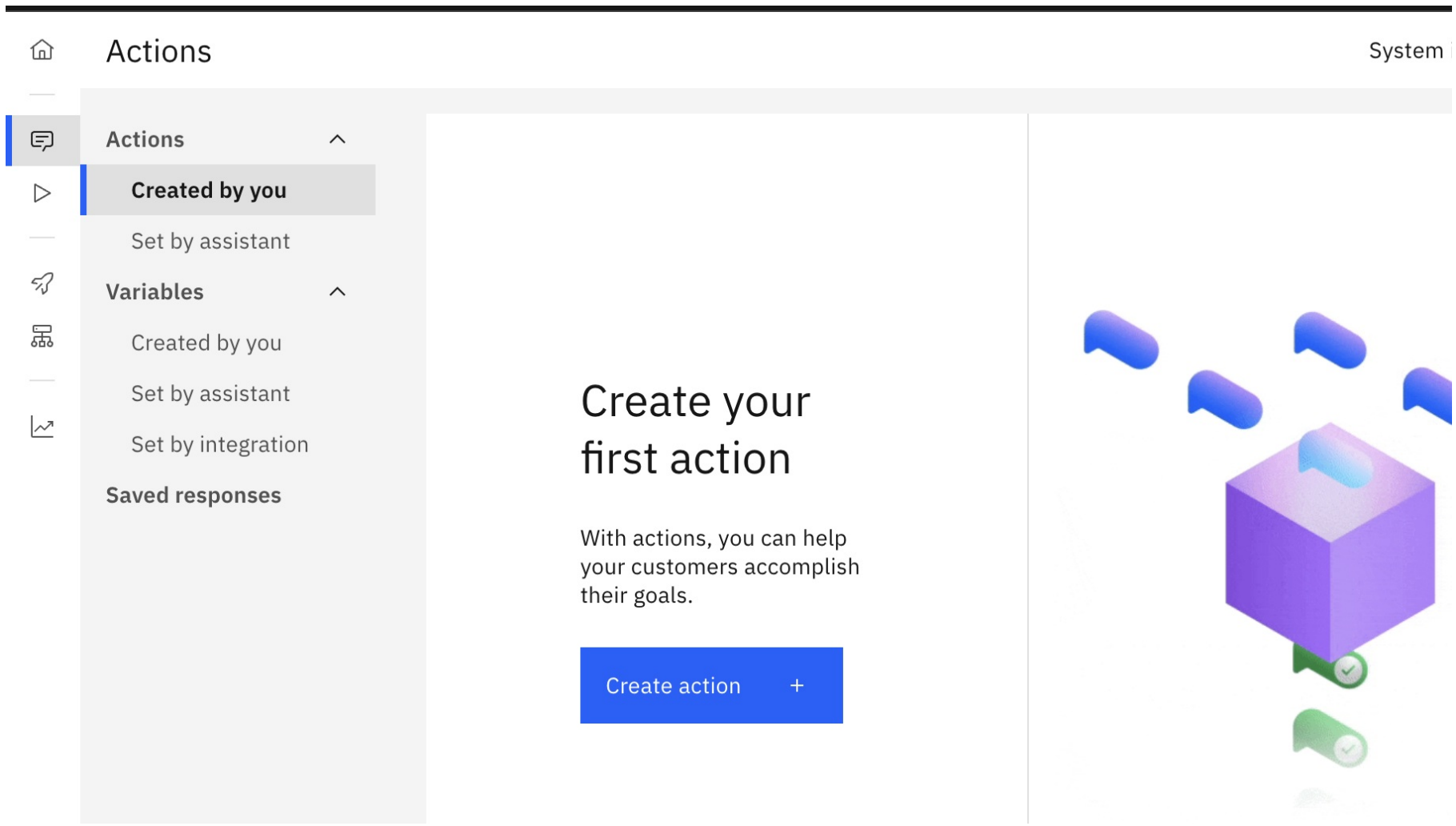

Figure 7. Actions page

2. Click **Create action**, and then choose to start from scratch.

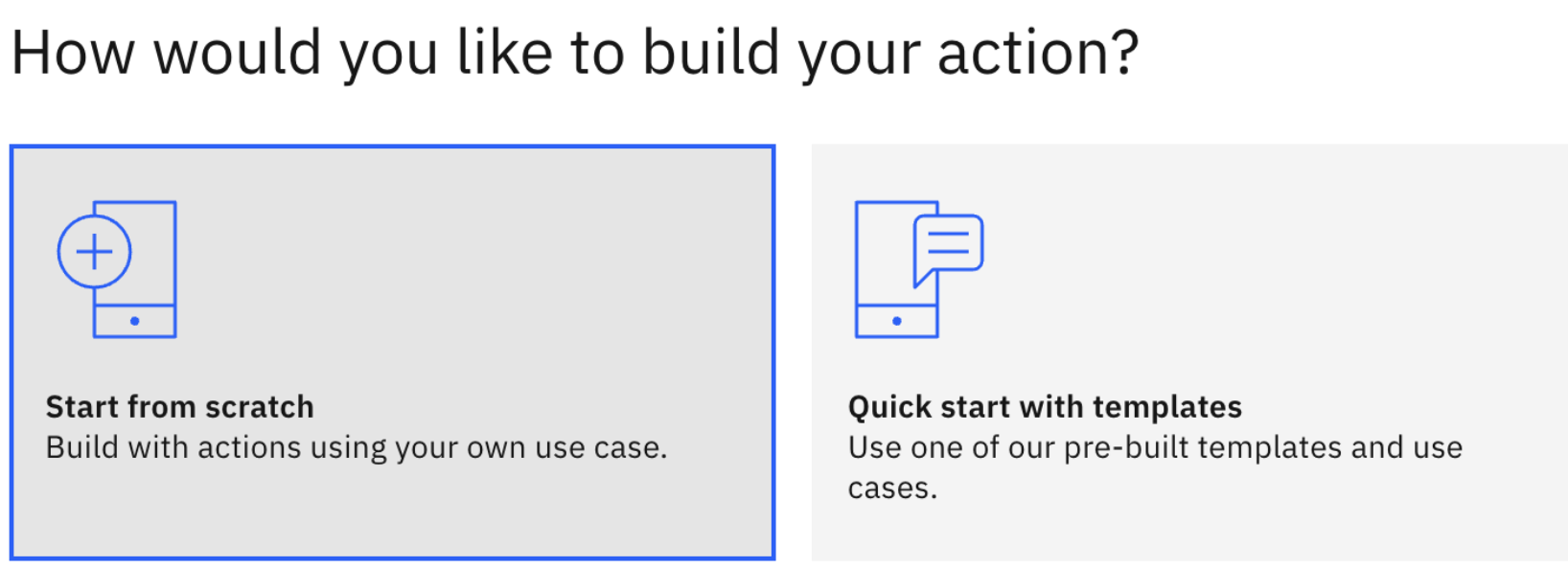

Figure 8. Action creation method options

3. Because you want the assistant to recognize when customers ask about economic research, add the following sample user question, and then click **Save**:

```
What are the latest working papers about?
```

   The editor closes. We want to add a few more examples.

4. Click the *Customer starts with* tile to continue adding examples.

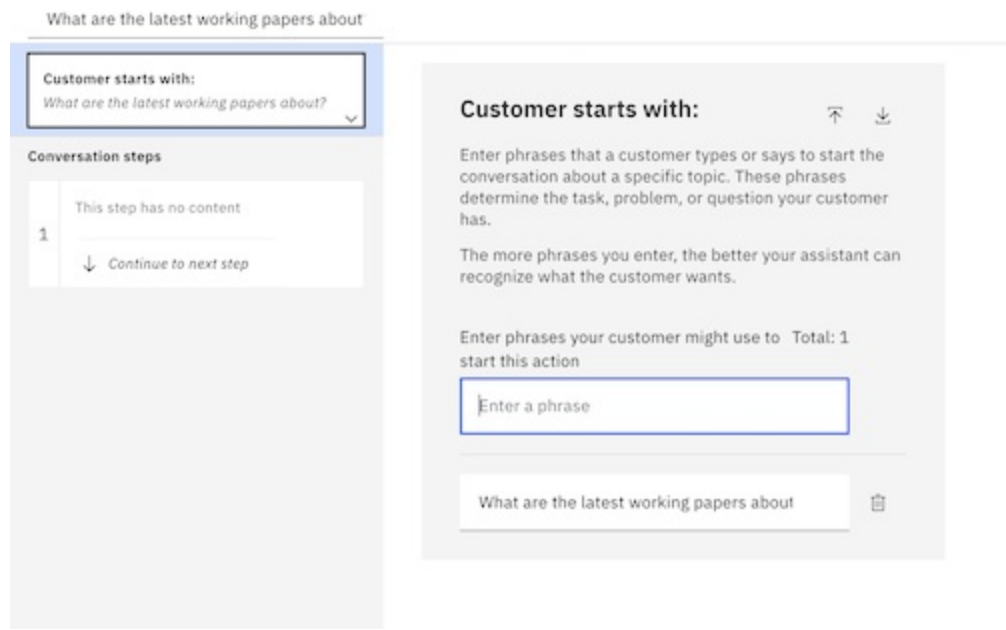

Figure 9. User question examples

5. Add the following questions:

```
Are there any working papers on the shipping industry?
```

```
Are there any papers that focus on inflation?
```

```
Are there papers about how trade policy affects pricing?
```

```
What's the latest research on municipal bond markets?
```

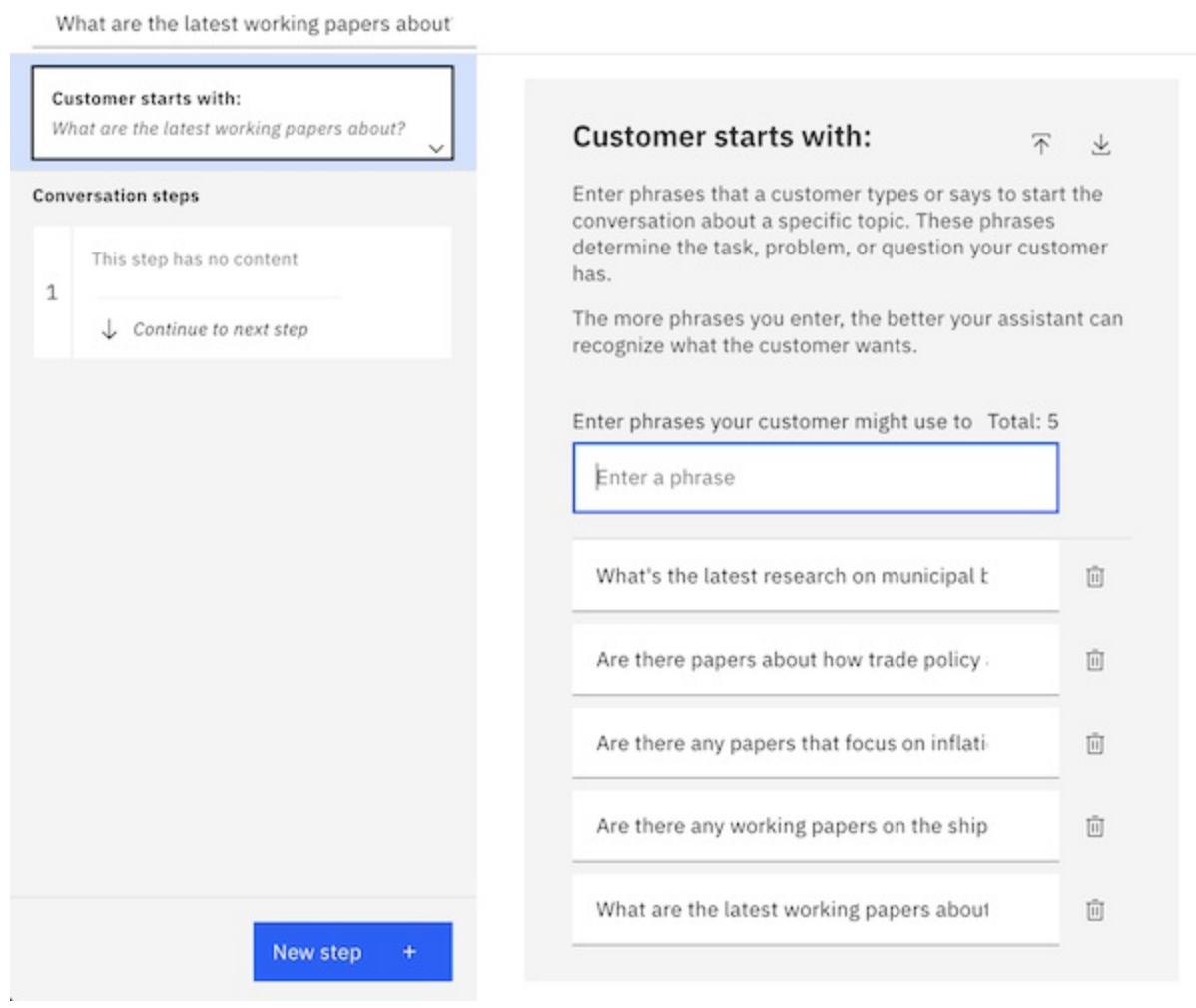

Figure 10. User examples list

6. Click the first step in the *Conversation steps* section.

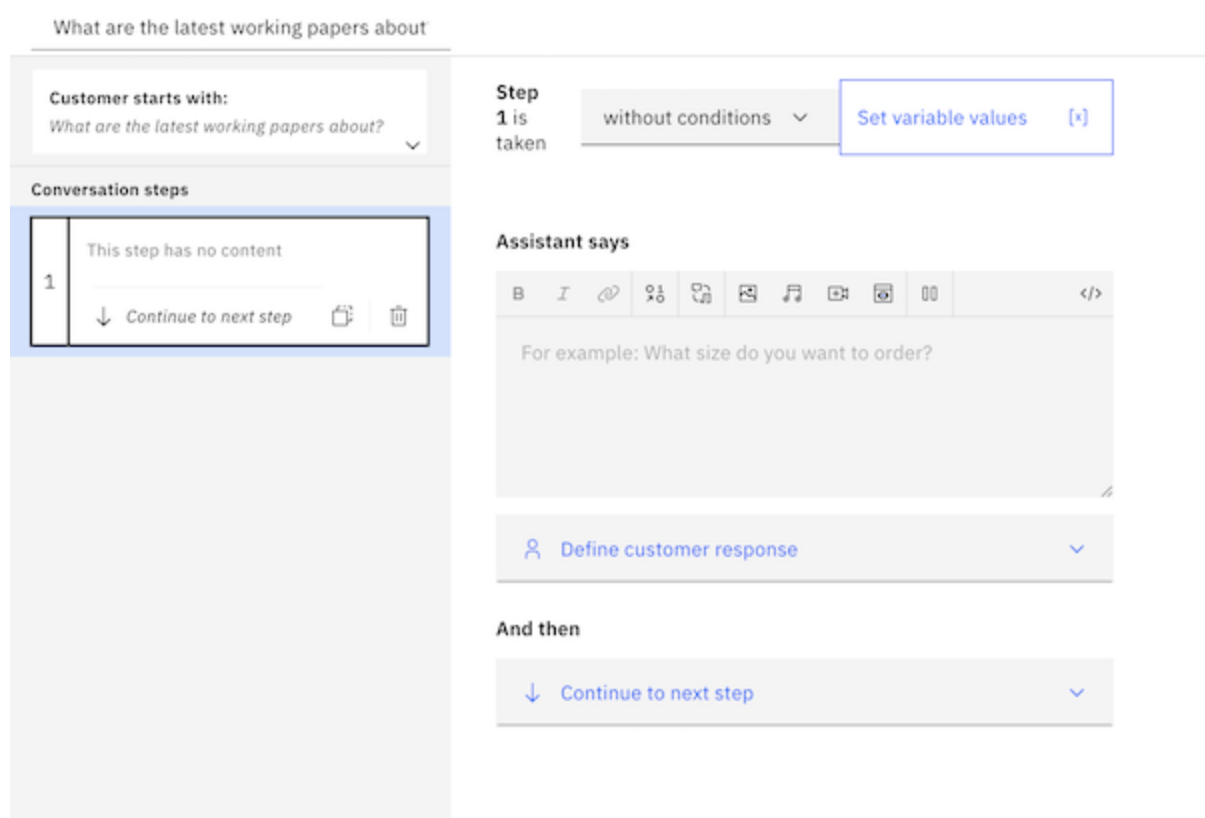

Figure 11. First step

7. Add the following text to the *Assistant says* field:

```
I'll check the Federal Reserve Economic Data website.
```

8. Do not add a customer response. Instead, in the *And then* section, click **Continue to next step**, and then choose **Search for the answer**.

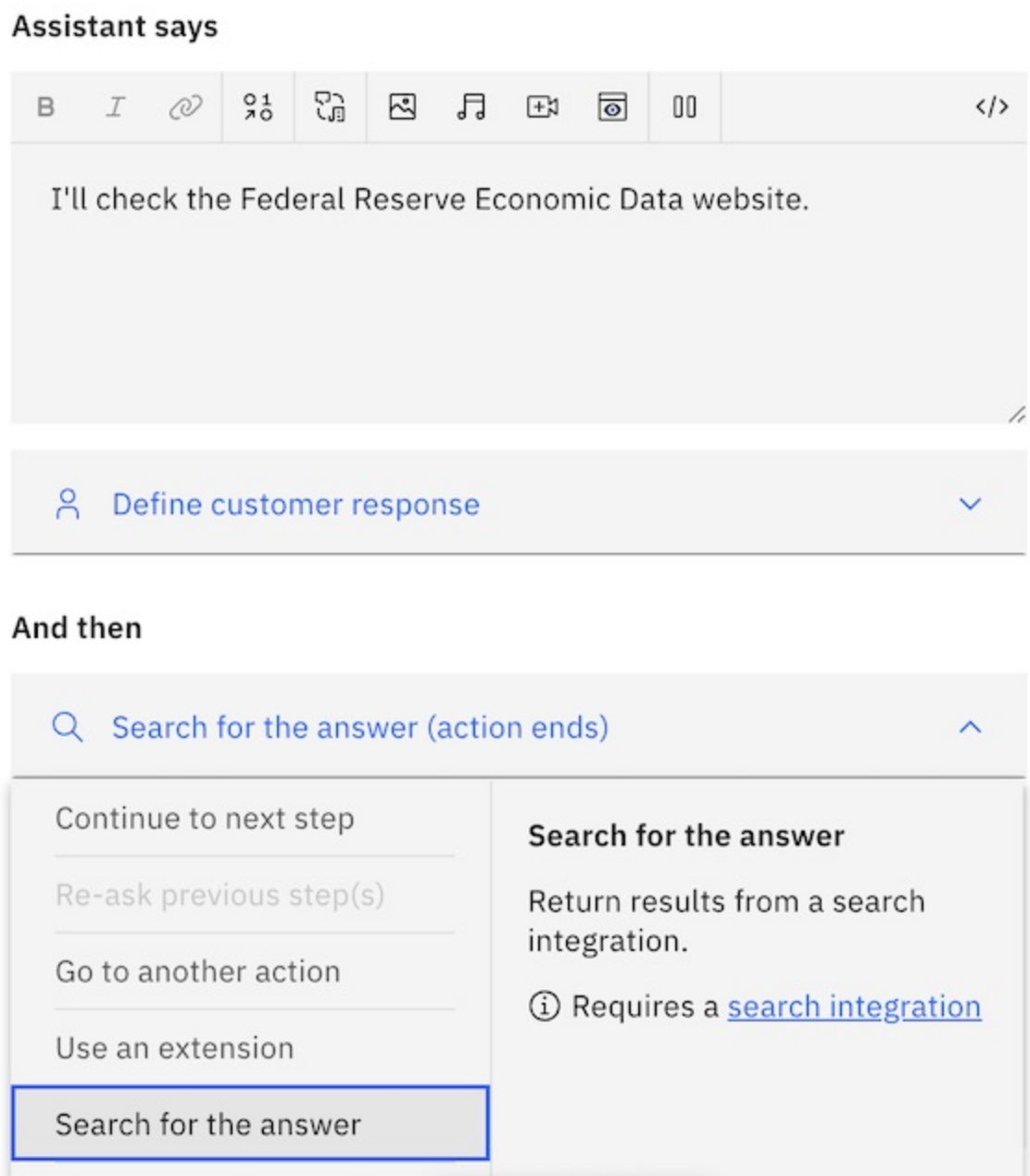

Figure 12. And then options

9. Click **Edit settings**.

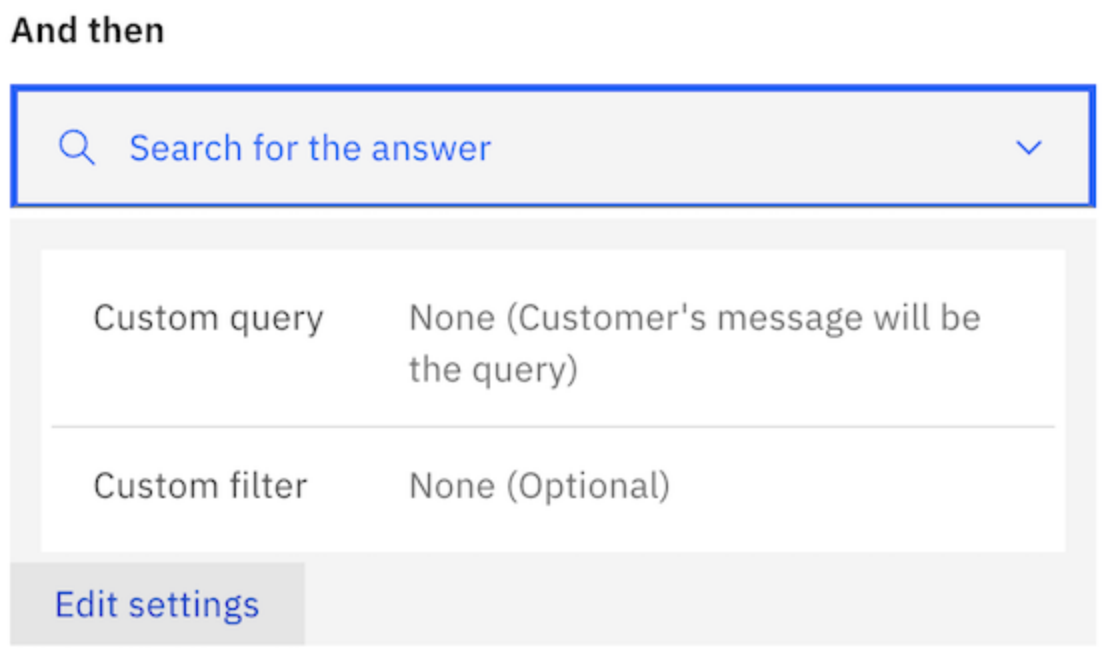

Figure 13. Search step

10. Select **End the action after returning results**, and then click **Apply**.

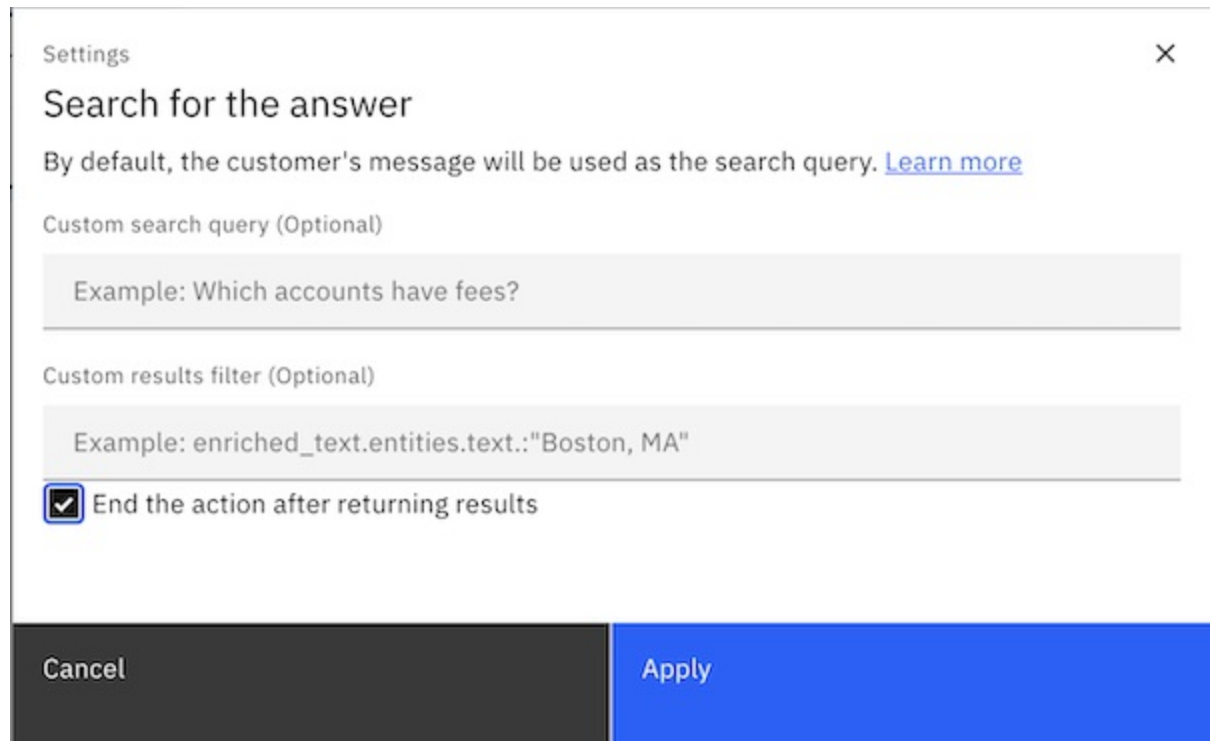

Figure 14. Search step settings

11. Save your changes, and then click the X to close the step.

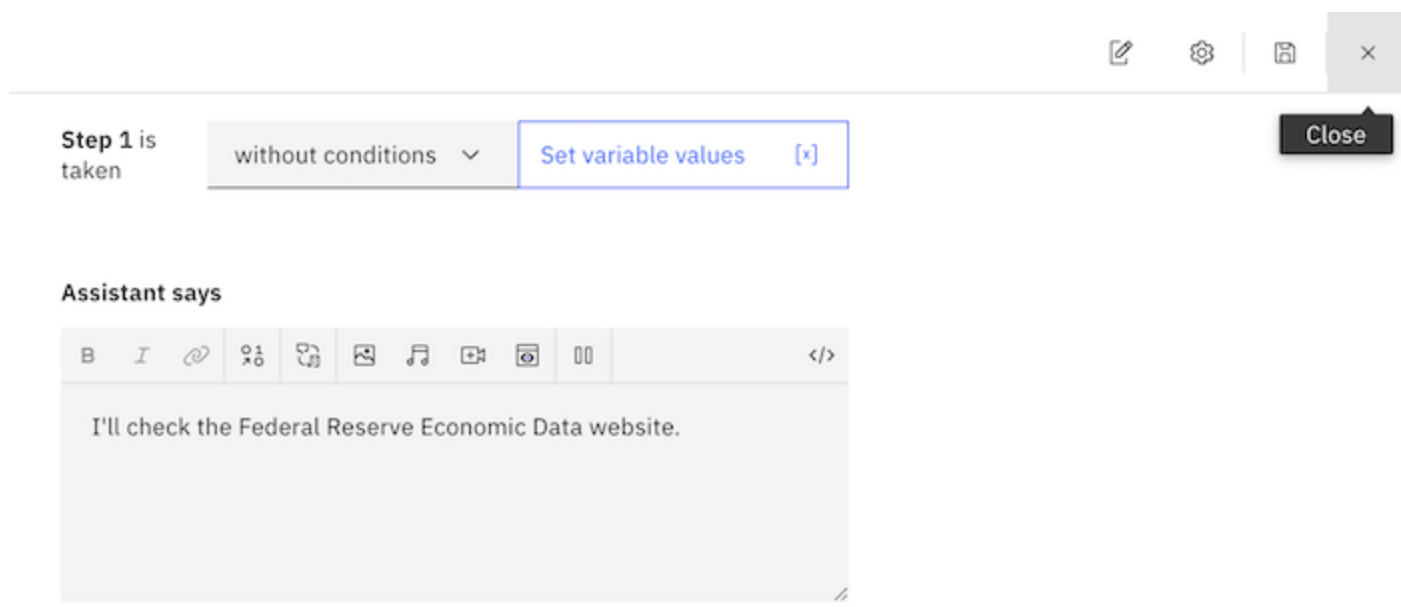

Figure 15. Close action

Congratulations! You successfully created an action that recognizes questions about FRED research papers and returns a search response.

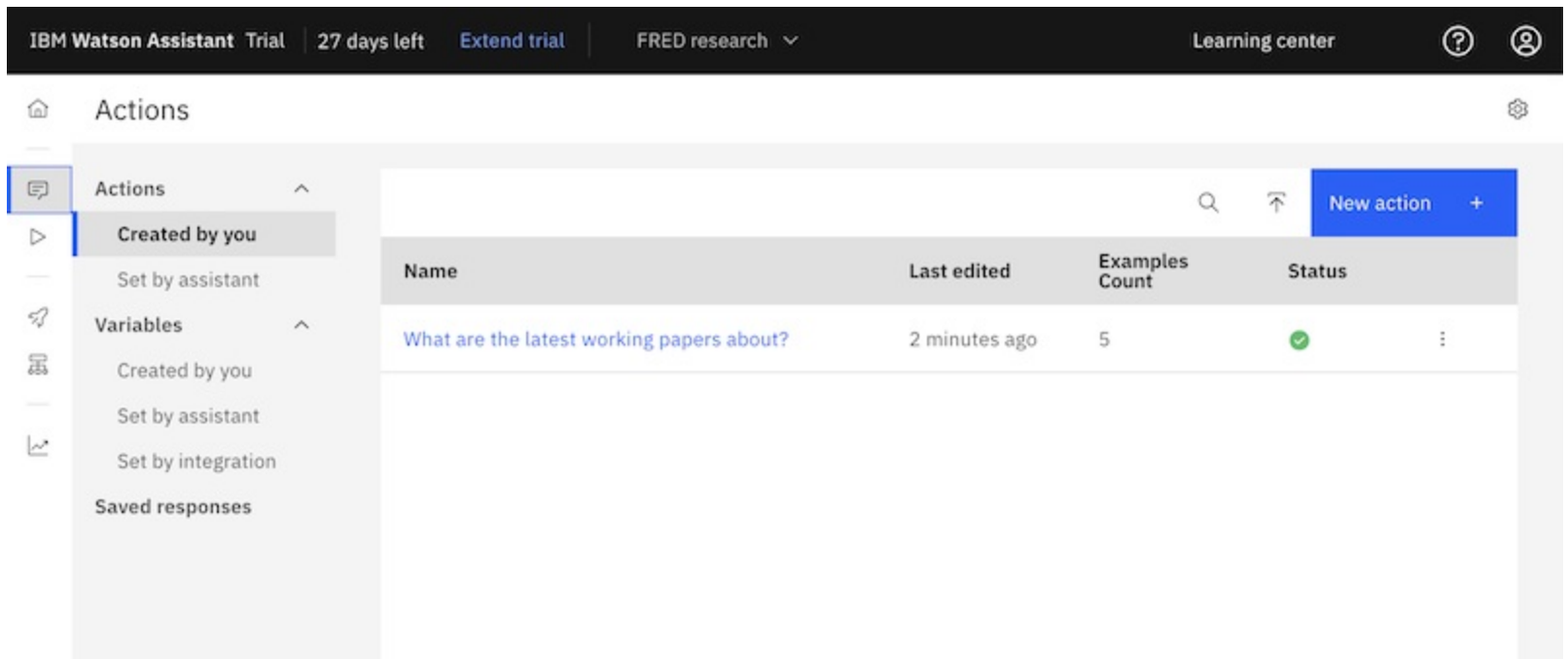

Figure 16. Created action

In a later step, we will connect the search response in this action to a search extension that is configured for the assistant.

Step 3: Create a Conversational Search project

Now that the assistant can recognize questions about a subject, let's give it access to data from which it can retrieve accurate answers.

In Discovery, create a Conversational Search project type. This project type is optimized for retrieving answers during dialog-driven interactions. For example, unlike other project types, it does not apply prebuilt enrichments that aren't needed.

1. Open a new web browser page.

   Tip: Keep the watsonx Assistant page open in a separate tab so you can switch between the two applications.

2. From the Discovery Plus plan service page in IBM Cloud, click **Launch Discovery**.
3. From the *My Projects* page, click **New Project**.
4. Name your project `Federal Reserve research`, and then click the **Conversational Search** tile.

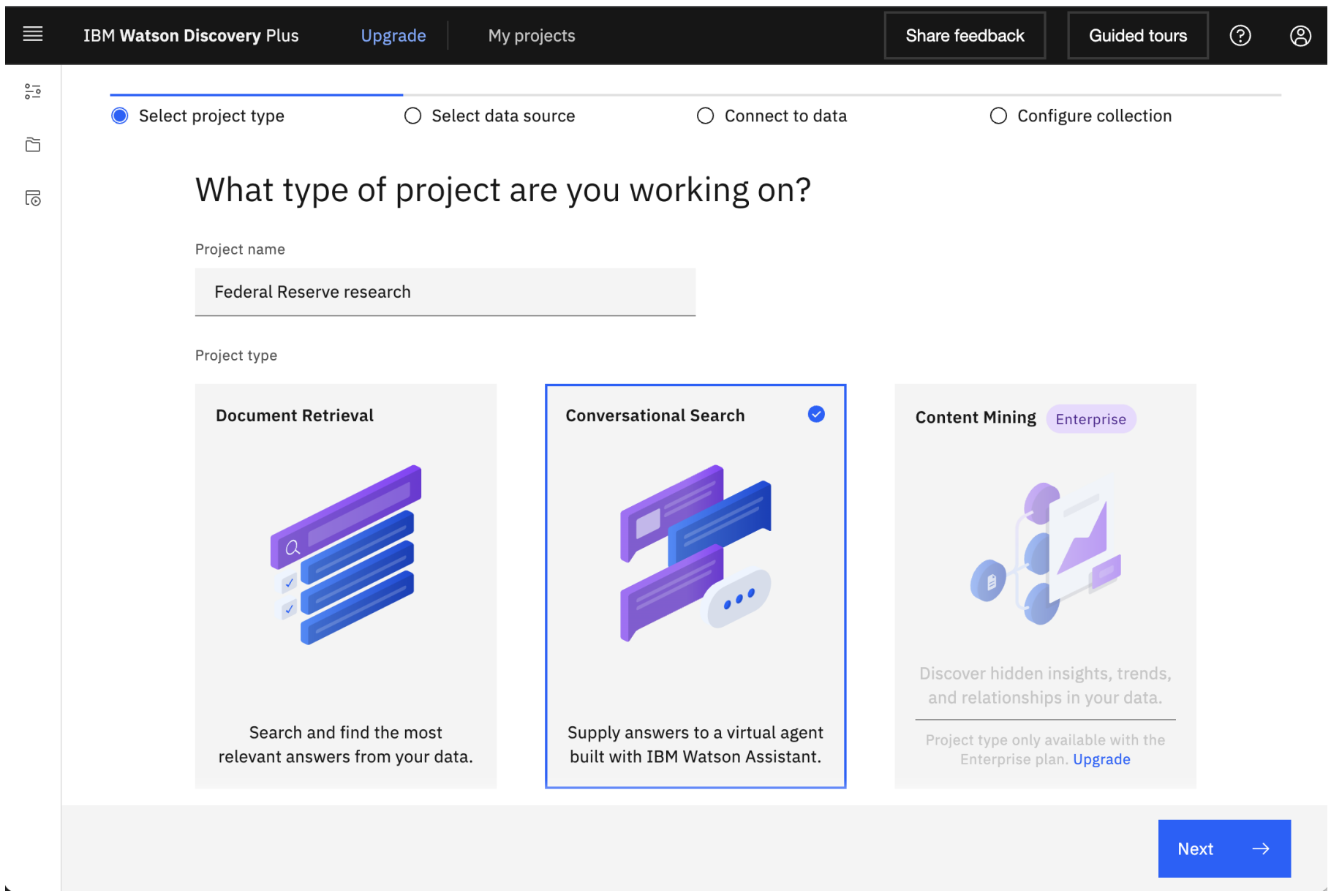

Figure 17. Project type options

5. Click **Next**.

You'll configure the data source for the project in the next step.

Step 4: Connect to a website

We want the virtual assistant to be able to answer questions about the latest working papers from the US Federal Reserve, so we will connect our project to the Federal Reserve Economic Data website that hosts the working papers.

1. From the *Select data source* page, click **Web crawl**, and then click **Next**.

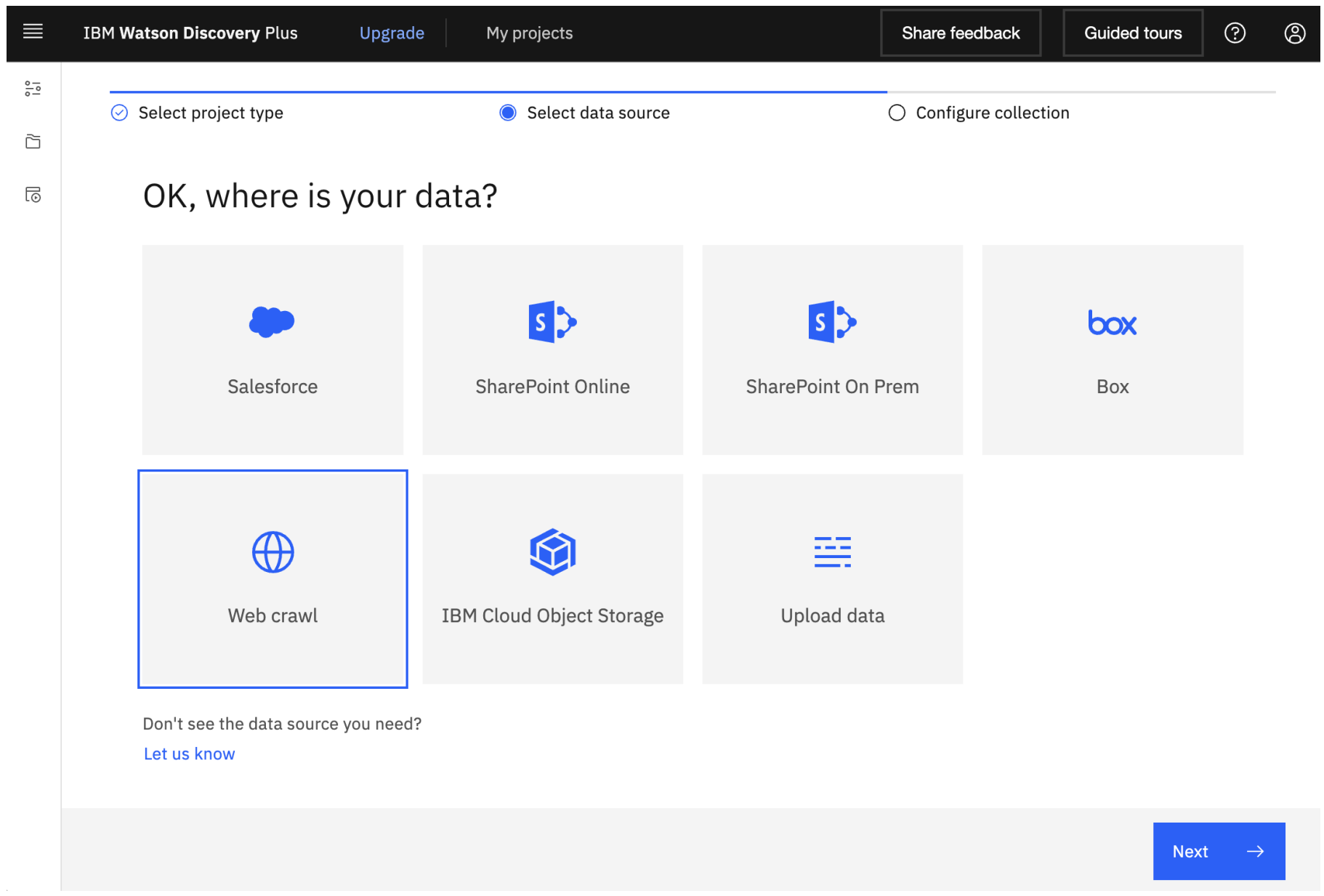

Figure 18. Data source options

2. In the **Collection name** field, add `FRED papers`.

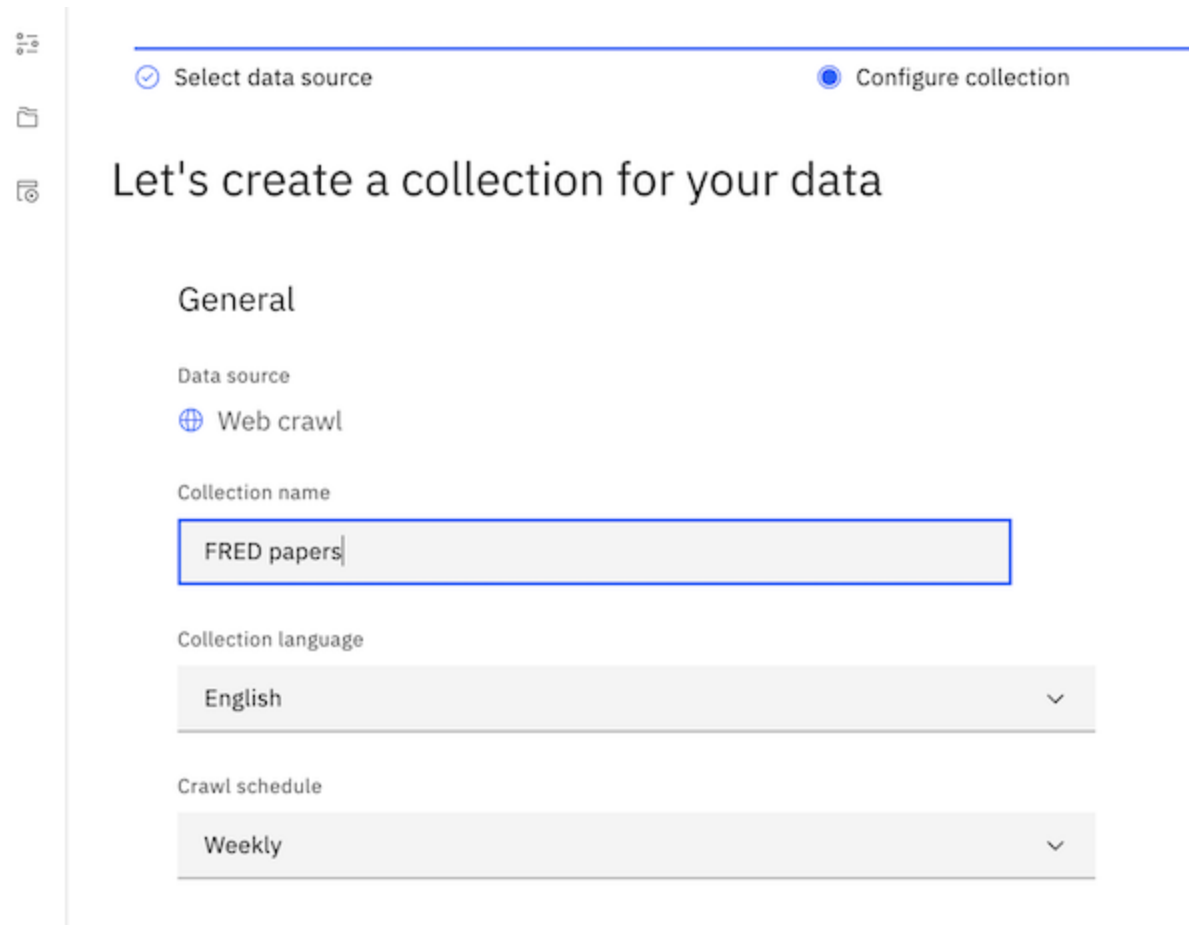

Figure 19. Web crawl connector

3. In the **Starting URLs** field, add the following URL:

```
https://research.stlouisfed.org/wp
```

You will add only one starting URL. In a real scenario, you might add multiple URLs that go to other pages with information about the same topic. By adding more URLs, you can expand the breadth of the expertise of your assistant.

4. Click **Add**.
5. Click the Edit icon for the URL that you just added.
6. In the **Maximum number of links to follow** field, change the value to 5.

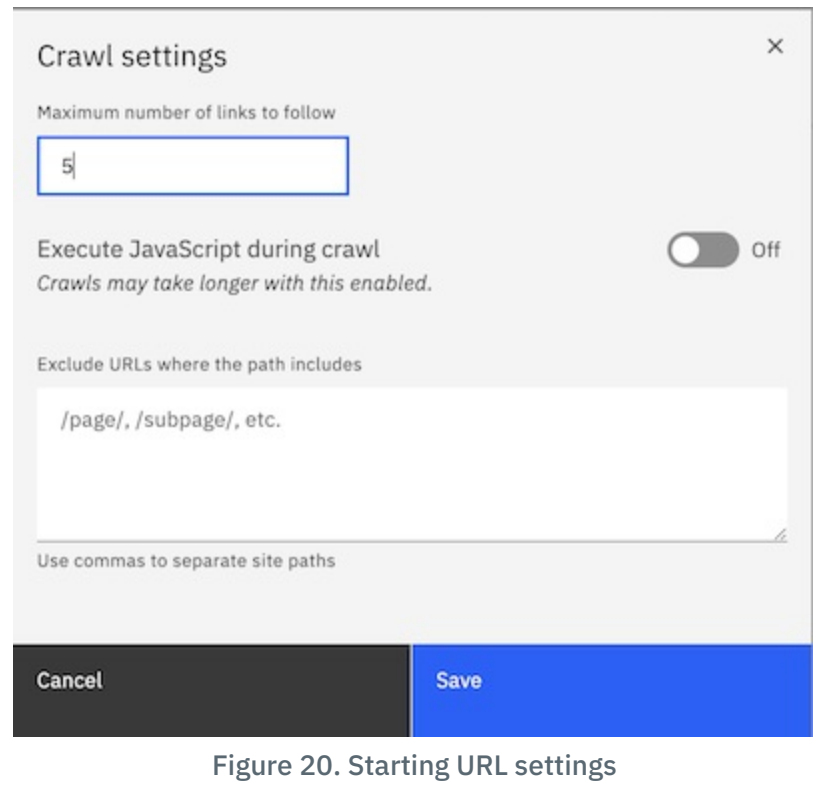

Figure 20. Starting URL settings

   By changing the value to 5, you indicate that you want the service to process the page that you specified plus you want it to follow up to 5 links from the starting page.

7. Click **Save**, and then click **Finish**.

The Discovery service crawls the web page that you specified starting with the page that you specified as the starting URL.

While the website is being crawled and the data indexed, let's go back to our watsonx Assistant service instance. It's time to connect the action that we created to this Discovery project.

Step 5: Add a search extension

Let's connect your assistant to your Discovery data.

1. From the navigation panel in watsonx Assistant, click **Environments**.

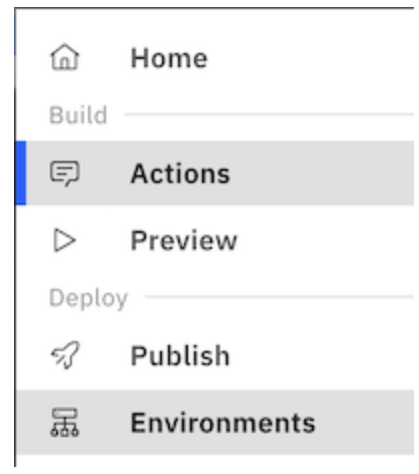

Figure 21. Environment menu

The draft environment is displayed. It shows that a web chat is connected to your assistant.

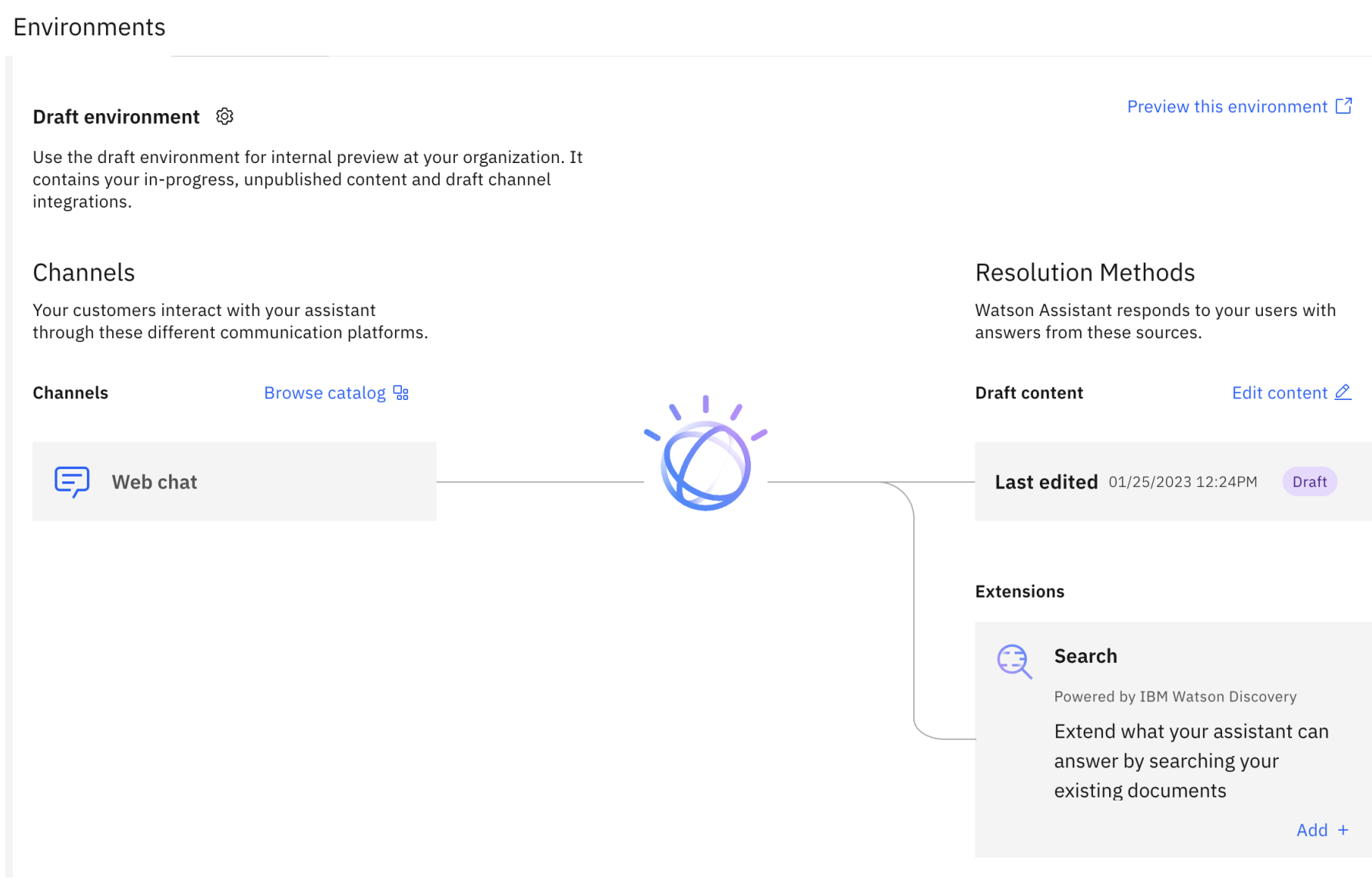

Figure 22. Draft environment diagram

2. Click the Web chat tile to edit the web chat.

   We don't want to add multiple starter questions, so we are going to turn off the home screen for the web chat. Click the **Home screen** tab.

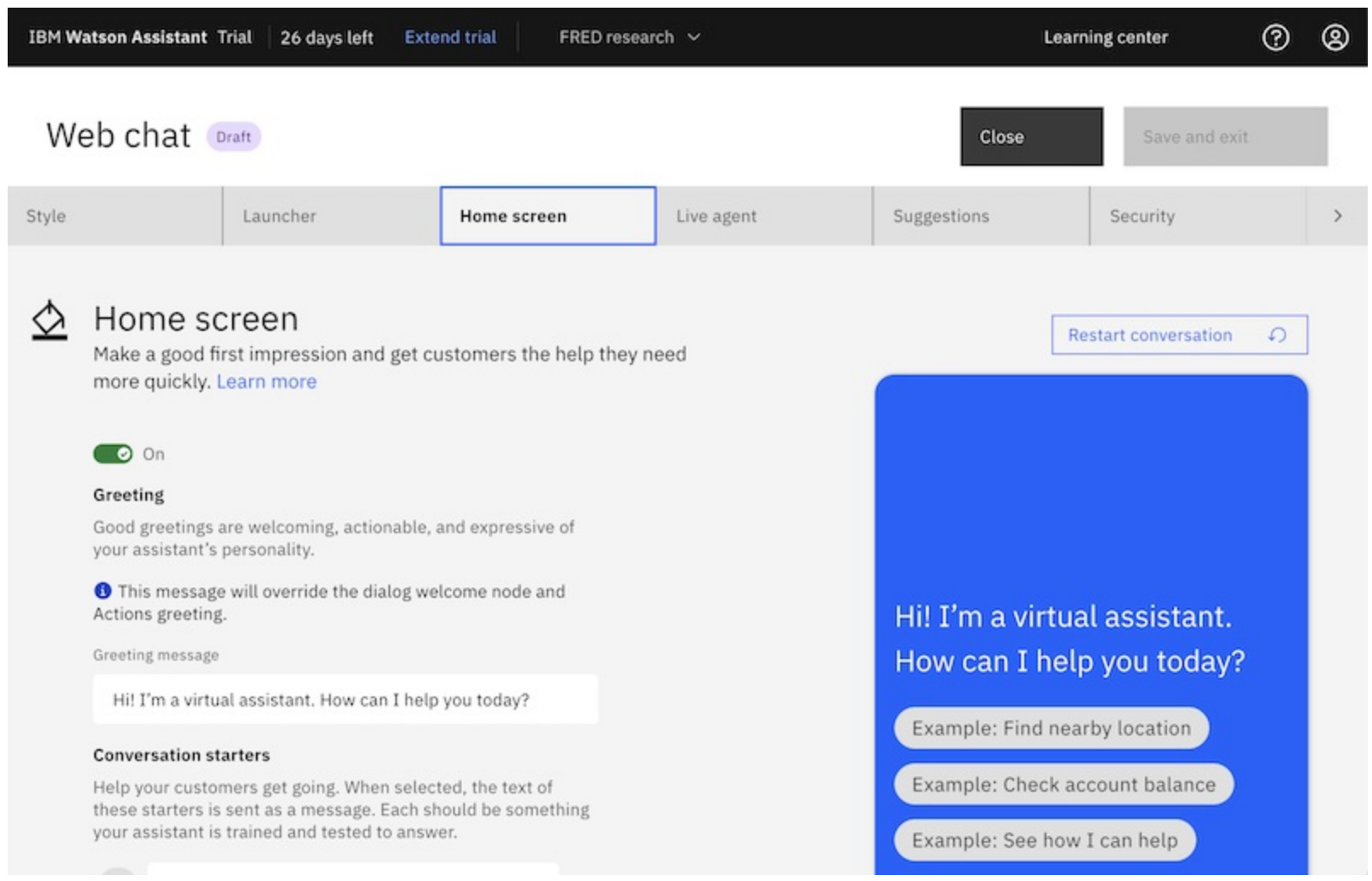

Figure 23. Web chat home screen configuration

3. Set the switcher to **Off**, and then click **Save and exit**.

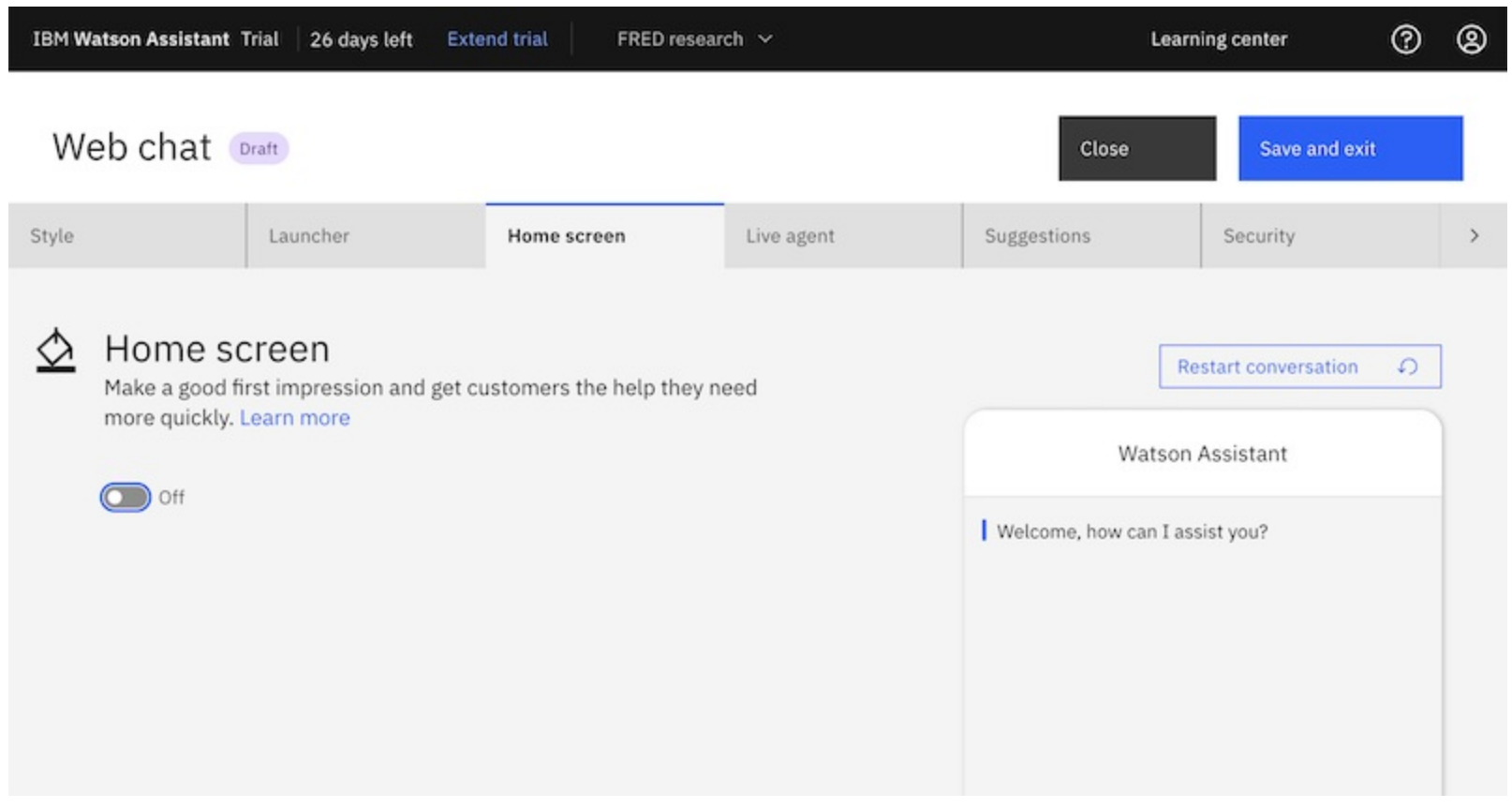

Figure 24. Web chat home screen disabled

4. We want to connect the web chat to a search extension. Click the **Add** button in the Search extension tile.

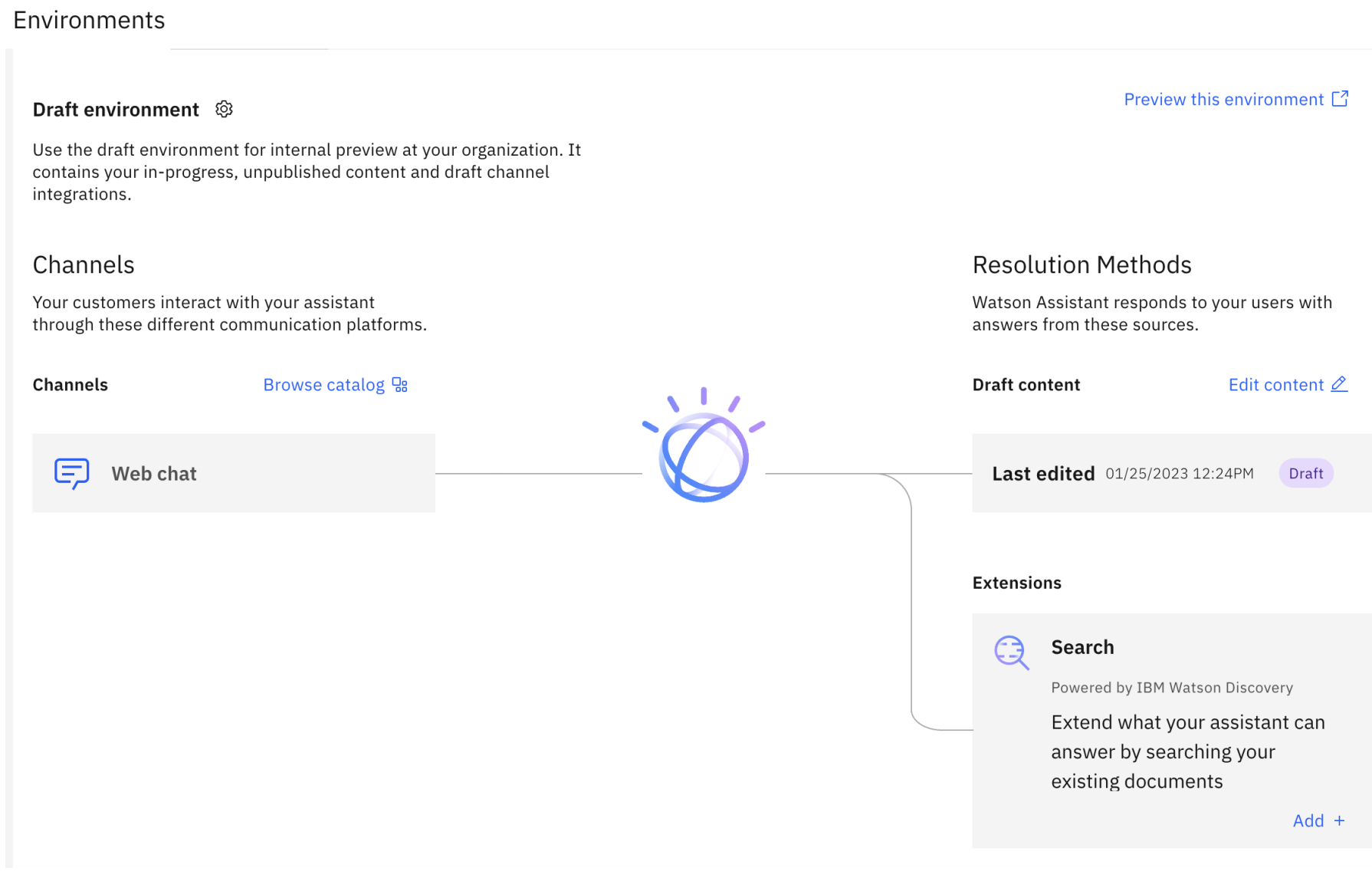

Figure 25. Search extension in draft environment

The Search Integration page is displayed.

5. Select the Discovery instance where your project is stored, and then select the **Federal Reserve research** project that you created earlier. Click **Next**.

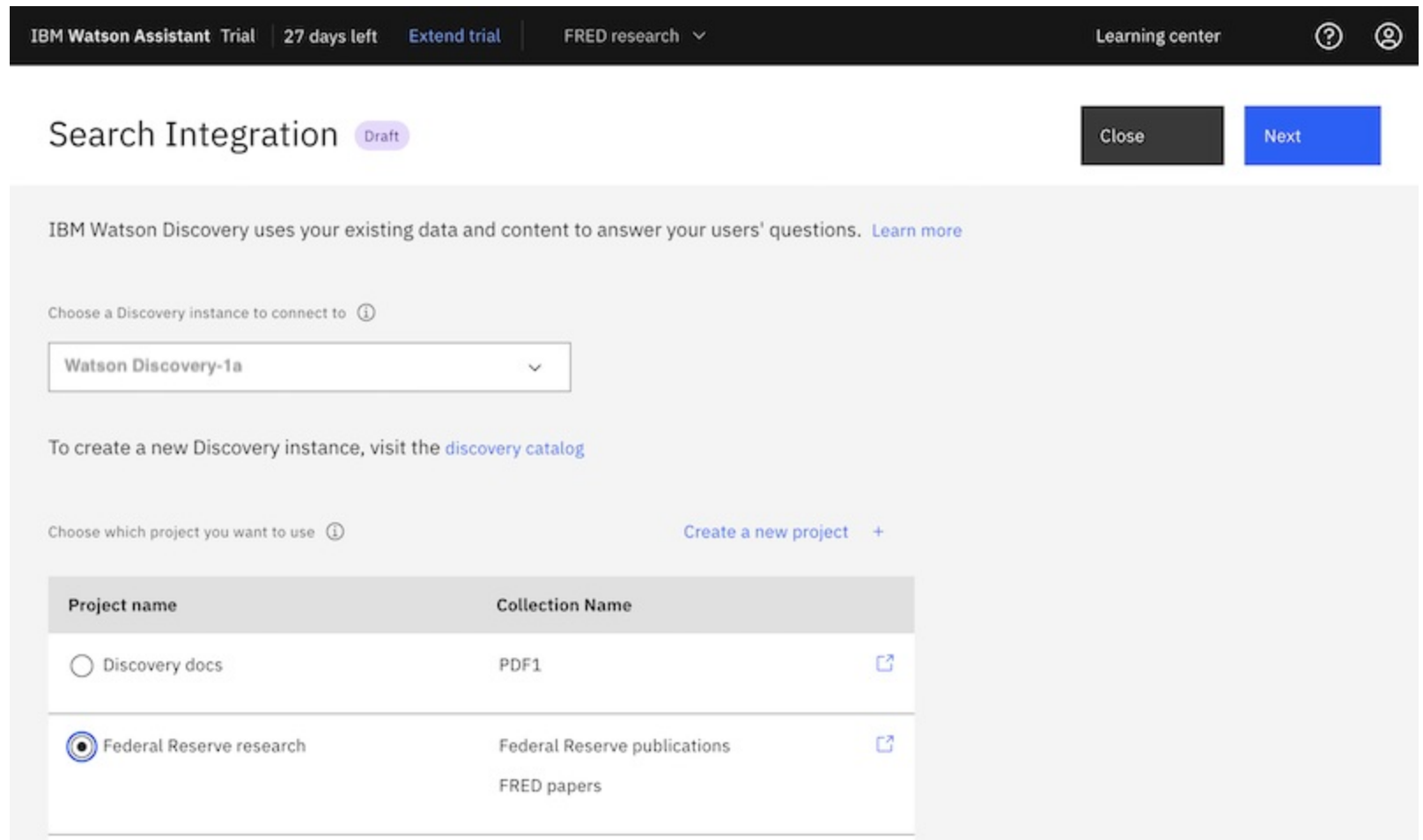

Figure 26. Search extension configuration

6. The default result content configuration uses the best fields; you don't need to change them.
7. In the *Define the text your search will display to the end user* section, edit the content to show the following message:

   ```
   The Federal Reserve Economic Data website has this information:
   ```

   Verify that the *Emphasize the answer* switch is set to **On**. This setting adds the `find_answers:true` parameter to the query request. As a result, a succinct answer to the query is shown in bold in the response that is returned by the assistant.

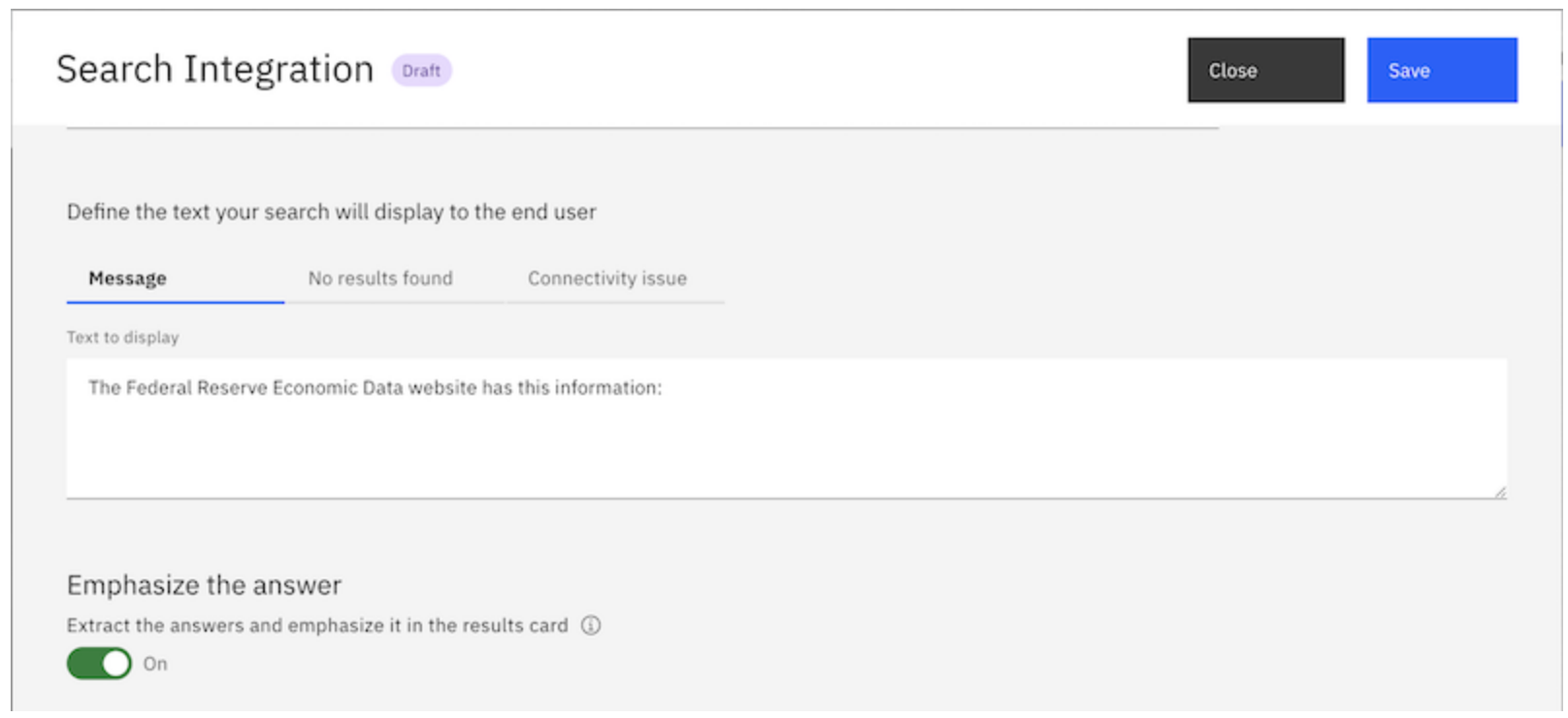

Figure 27. Search extension settings configured

8. Click **Create**.

Step 6: Preview the assistant

To preview an assistant that connects to data that is stored in Discovery, you must preview the assistant from the Environments page. When you test it separately, the assistant is not able to retrieve data from Discovery.

1. From the Environments page, click **Preview this environment**.

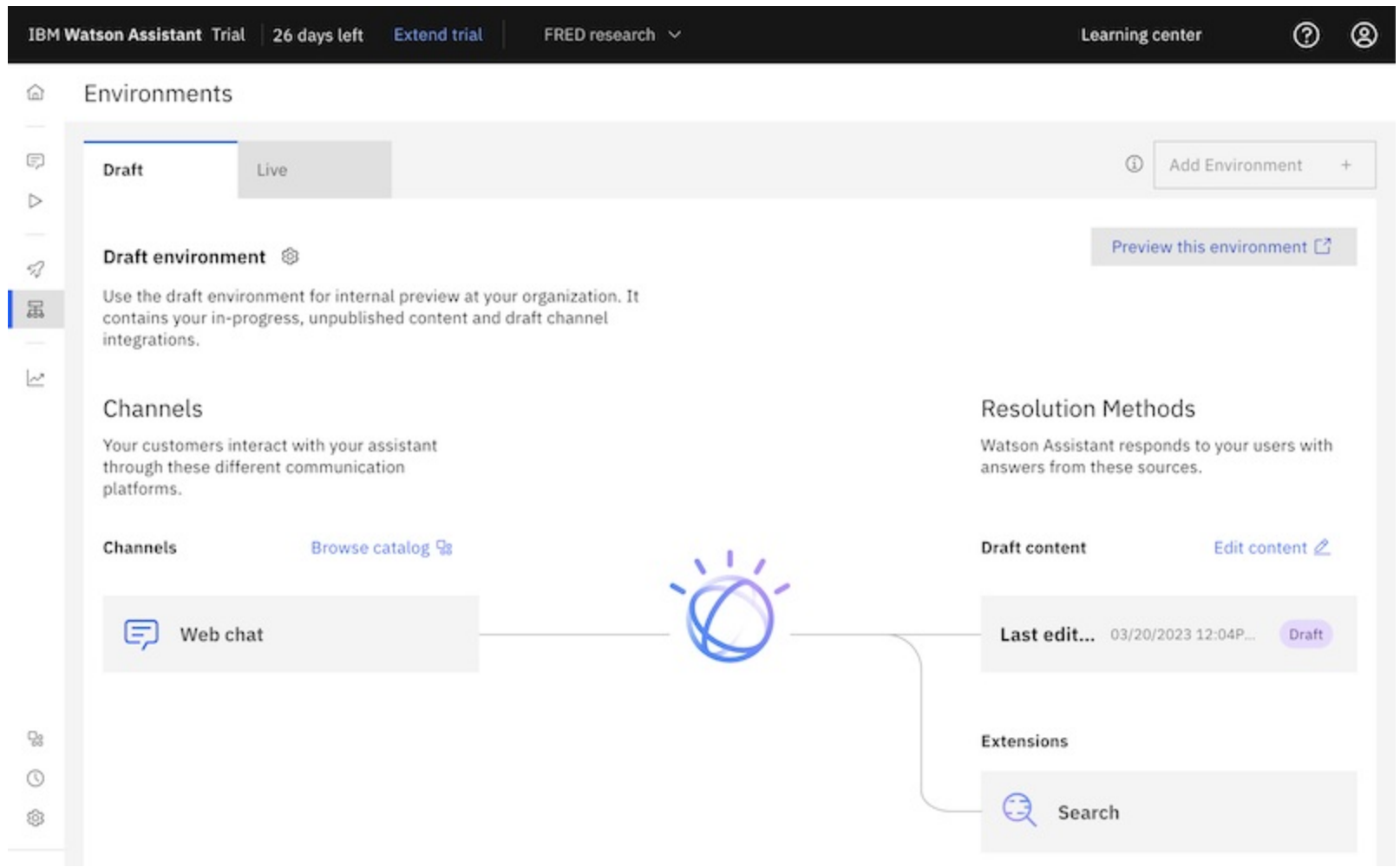

Figure 28. Search extension enabled

A sample web page is displayed that includes a chat icon.

Figure 29. Web chat icon

2. Click the chat icon to open the web chat window.

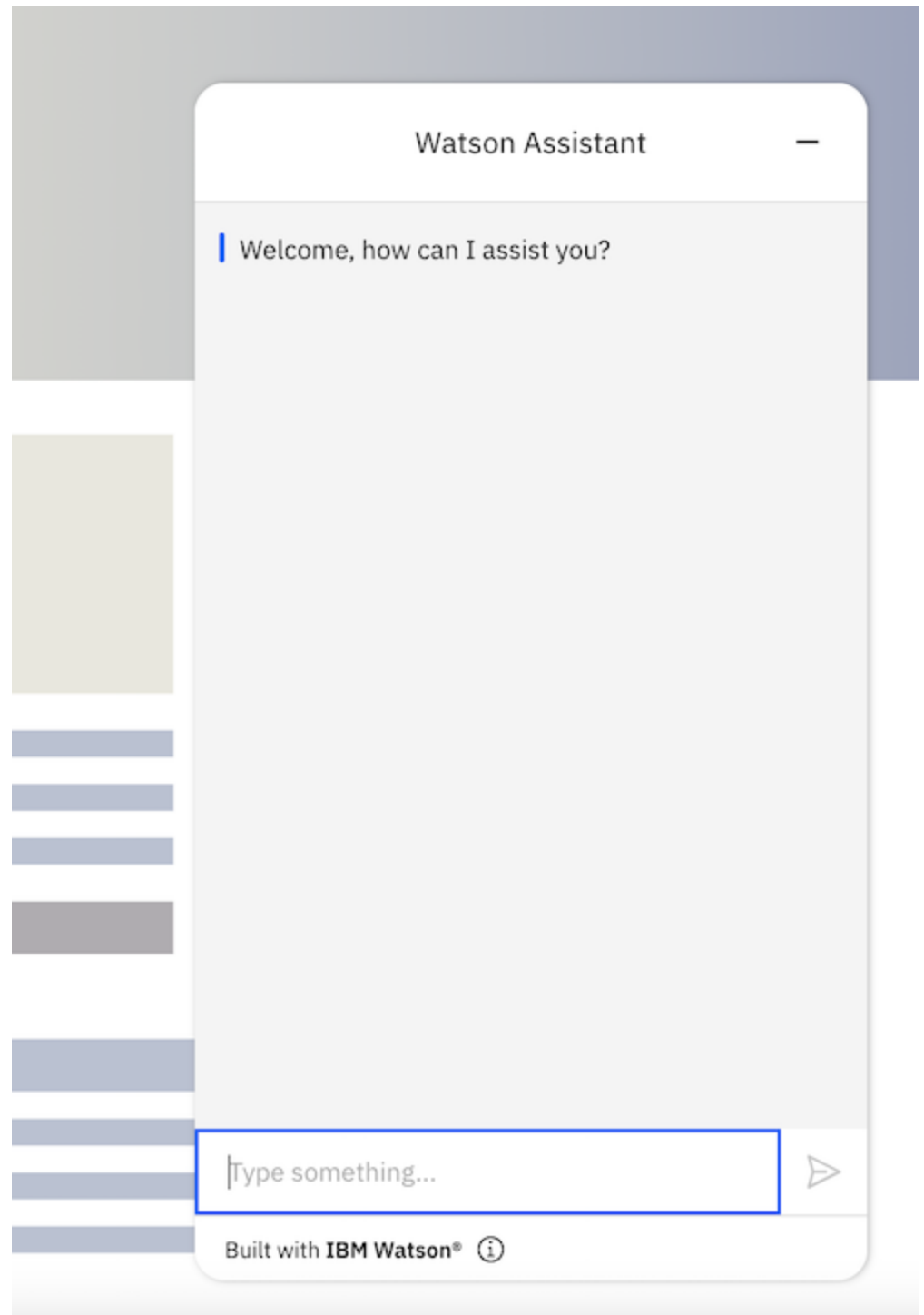

Figure 30. Web chat welcome message

3. Enter the following text question:

```
What impact is inflation having on the real estate market?
```

> **Note:** This test question is not one of the questions that we used to train the assistant.

The correct answer is returned and it includes a link to the source documentation page.

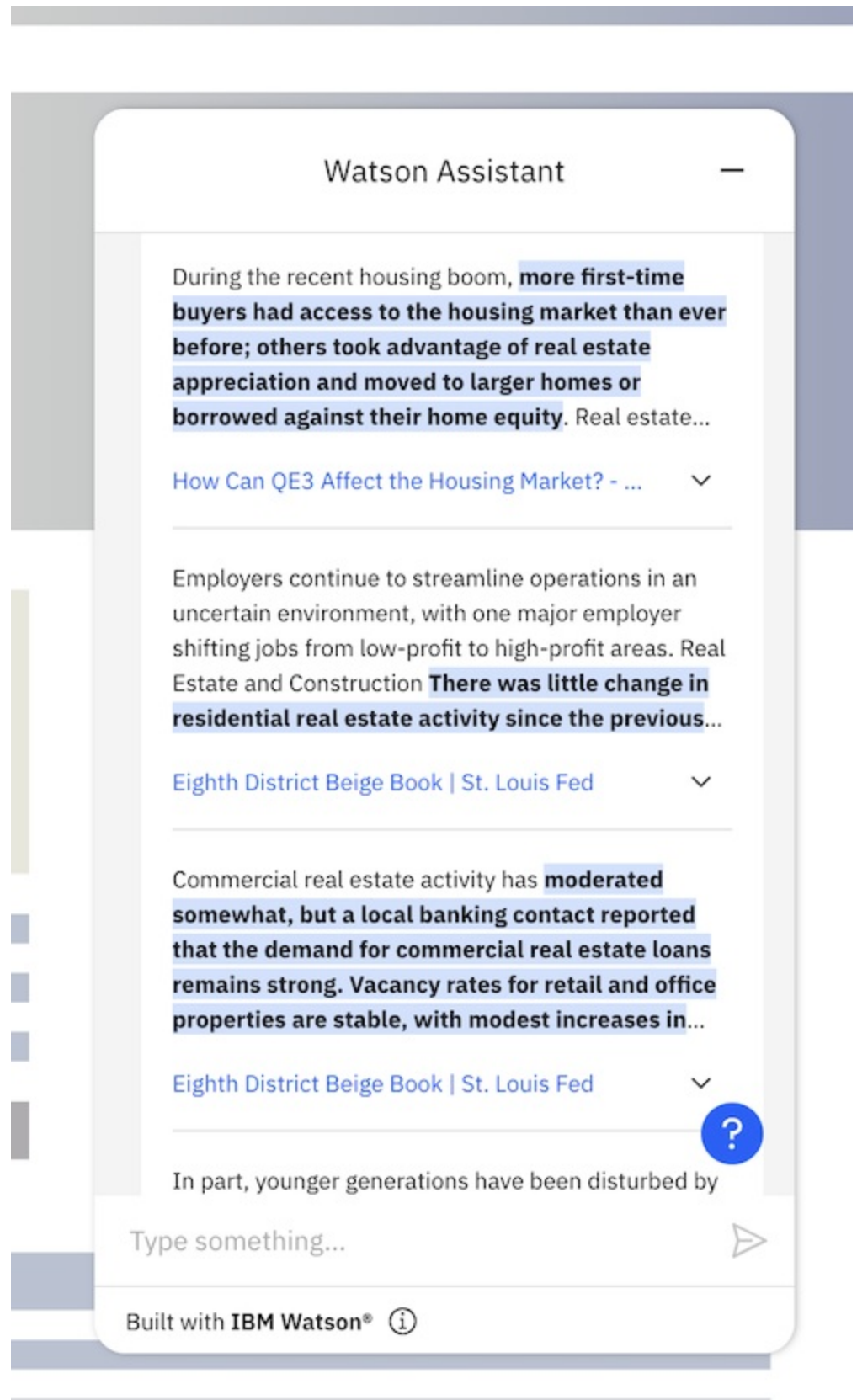

Figure 31. Web chat returns search response

Congratulations! You successfully created an assistant that can answer questions about economic topics by retrieving information from working papers that are available from the US Federal Research Economic Data website.

Summary

In this tutorial, you created a Watson Discovery Conversational Search project with a web crawl connector that collects information about working papers from the US Federal Reserve Economic Data website. Separately, you created a watsonx Assistant virtual assistant with a single action that can recognize user questions about economic subjects. You added a Search extension to your assistant that connects the action's search response to the Discovery project where economic data is stored. Finally, you tested your virtual assistant by asking a question and getting a useful response that featured data from relevant economic research papers.

Next steps

The assistant that you created and connected to a search extension is available from the Draft environment. Next, you can publish your assistant to a production environment and deploy it. There are a variety of methods you can use to deploy the assistant. For more information, see Overview: Previewing and publishing.

Use Smart Document Understanding (SDU) to improve search results

In this tutorial, you use the Smart Document Understanding feature of the Discovery service to create a user-trained Smart Document Understanding (SDU) model. You then split a single document into many smaller documents so that some types of answers are easier to find.

Note: This tutorial works with both managed and installed deployments.

Learning objectives

By the time you finish the tutorial, you will understand how to:

- Create a Document Retrieval project in Discovery.
- Upload a PDF document to your Discovery project.
- Use the Smart Document Understanding (SDU) tool to create a user-trained SDU model.
- Split a document into smaller, more consumable chunks.

Duration

This tutorial takes approximately 3 hours to complete.

Prerequisite

1. Before you begin, you must set up a paid account with IBM Cloud.

   You can complete this tutorial at no cost by using a Plus plan, which offers a 30-day trial at no cost. However, to create a Plus plan instance of the service, you must have a paid account (where you provide credit card details). For more information about creating a paid account, see Upgrading your account.

2. Create a Plus plan Discovery service instance.

   Go to the Discovery resource page in the IBM Cloud catalog and create a Plus plan service instance.

Important: If you decide to stop using the Plus plan and don't want to pay for it, delete the Plus plan service instance before the 30-day trial period ends.

Step 1: Create the Document Retrieval project

Create a project. Choose to create a Document Retrieval project type. This type is optimized for finding answers that are returned as passages from large documents.

For more information about project types, see Creating projects.

1. From the Discovery Plus plan service page in IBM Cloud, click **Launch Discovery**.
2. From the *My Projects* page, click **New Project**.
3. Name your project `Finance tutorial project`, and then click the **Document Retrieval** tile.

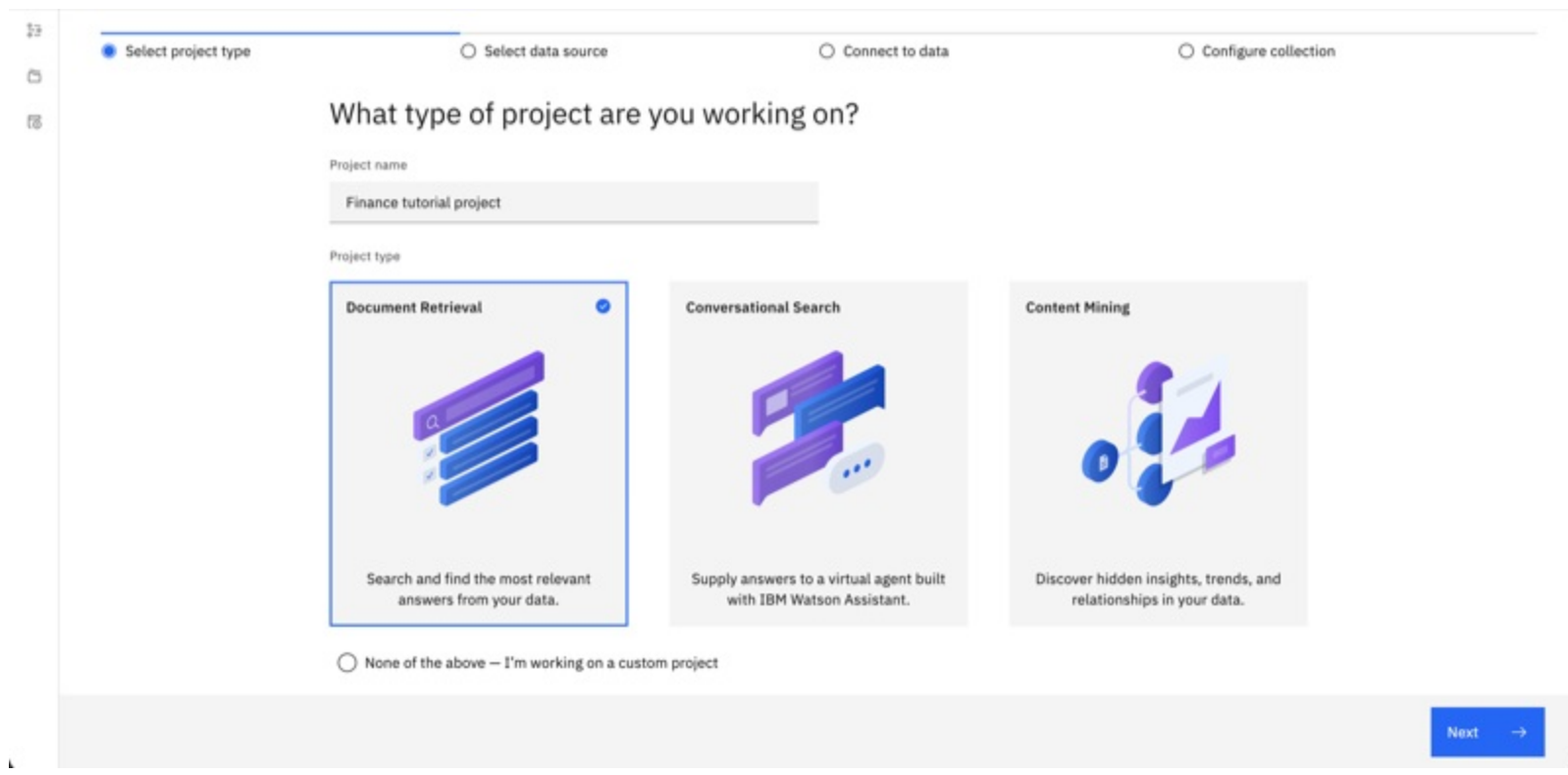

Figure 1. Project type options

4. Click **Next**.

You'll configure the data source for the project in the next step.

Step 2: Upload a PDF file

We want the search application to be able to answer questions about algorithmic trading. Therefore, we are adding the “Staff Report on Algorithmic Trading in US Capital Markets” PDF that was created on 5 August 2020 as a data source for the project.

1. Get a copy of the PDF so that you can upload it to your project. You can download the file from the US Securities and Exchange Commission website.
2. From the *Select data source* page, click **Upload data**, and then click **Next**.

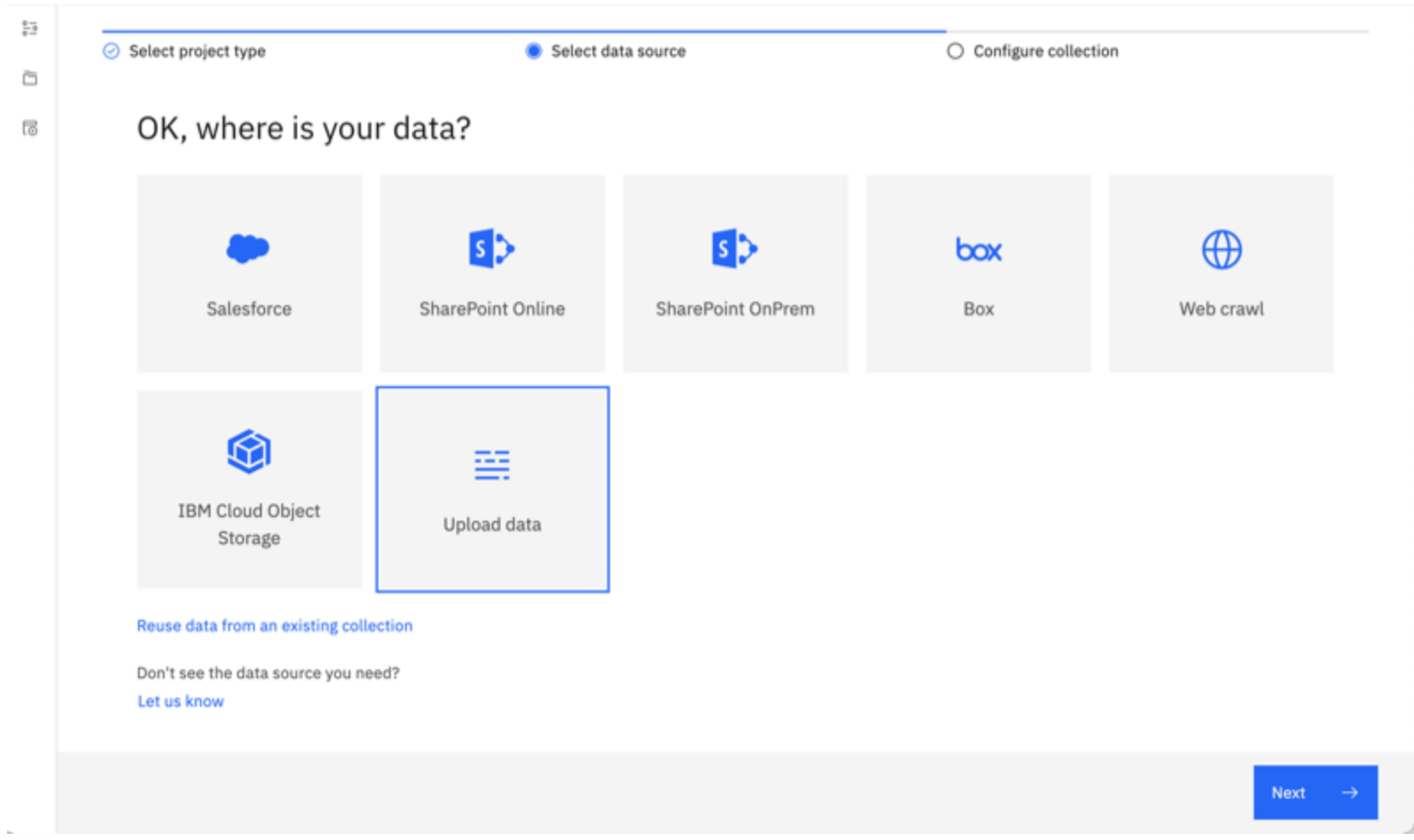

Figure 2. Data source options

3. In the **Collection name** field, add `Algorithmic Trading PDF`, and then click **Next**.

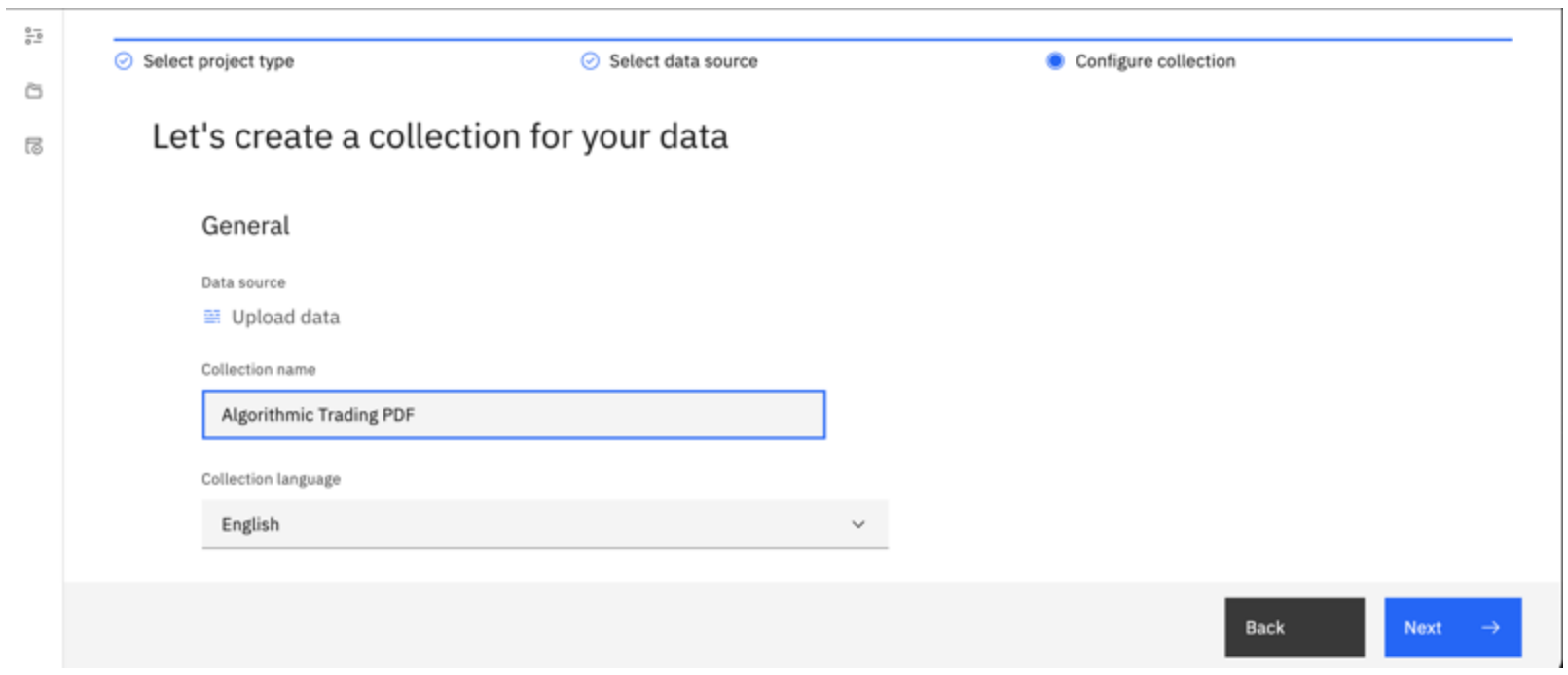

Figure 3. Uploaded data collection name field

4. Drag the file that you downloaded to the page and drop it into the tile with the *Drag and drop files here or upload* link.

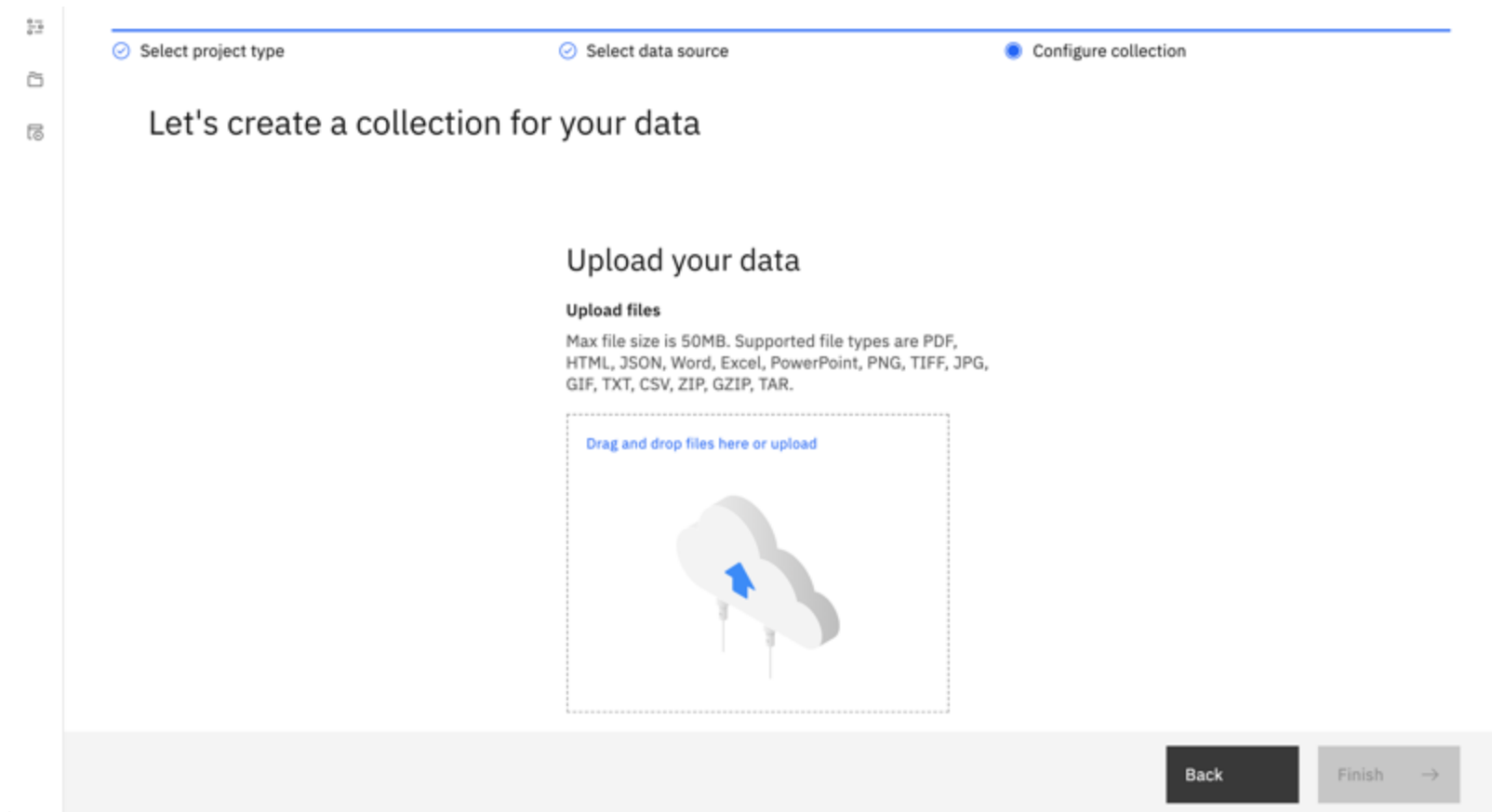

Figure 4. File upload dialog

5. Click **Finish**.

   You add only one file. In a real scenario, you might upload multiple files with information about the same topic. By adding more files, you can expand the breadth of the information that your search application can leverage.

The service uploads the document. As it uploads the document, Discovery crawls the data and indexes key information. Because you created a Document Retrieval project type, Discovery makes a note of the *Entities* information that it finds and recognizes as it crawls the document.

Step 3: Review the document

Analyzing and indexing the document can take a few minutes. While the processing is under way, review the source document to get a feel for its content. It is a good idea to understand the structure of your own documents before you use the Smart Document Understanding tool to annotate them.

Smart Document Understanding (SDU) uses visual imaging technologies to understand the structure of a document by analyzing the format and positioning of the text. You label sections of the document, such as subtitles or tables, to teach Discovery to recognize the sections. You can also label sections that you want the search function to ignore. For example, you might not want to search page footers or the table of contents information. After you teach the SDU tool to recognize footers, for example, you can exclude the footer field from the index.

1. Monitor the progress of collection processing by opening the *Activity* tab.

   Click **Manage collections** from the navigation panel.

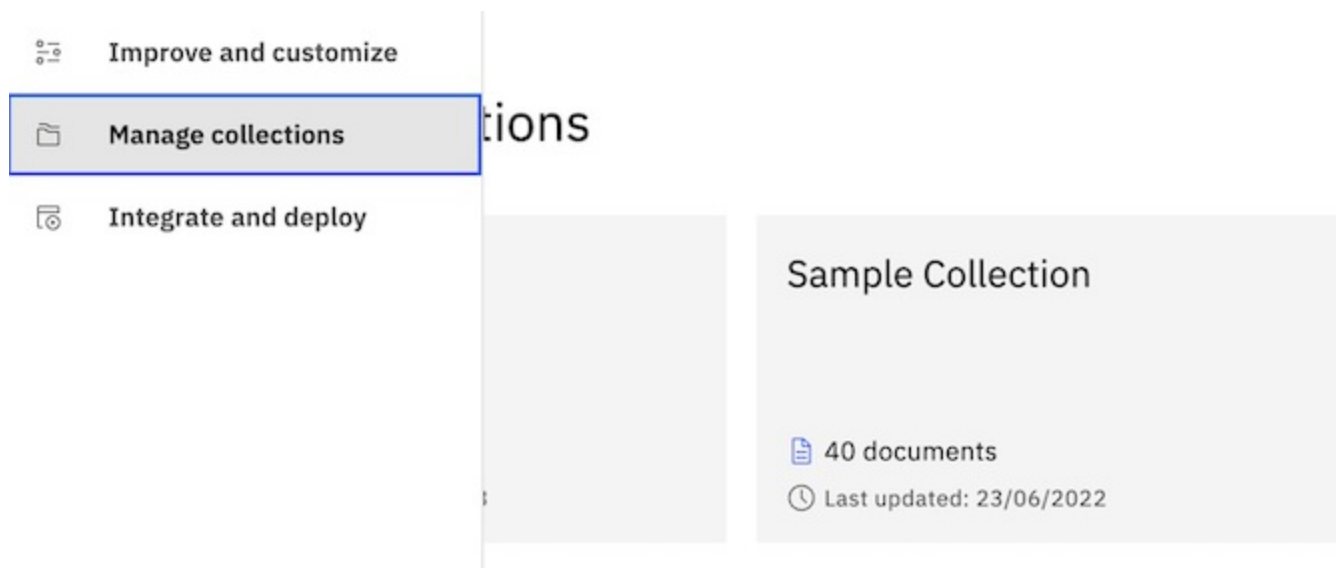

Figure 5. Manage collections menu option from the navigation panel

   Click the *Algorithmic Trading PDF* collection tile. The collection opens to the Activity tab.

2. While you wait for the collection to be ready, open the `Algo_Trading_Report_2020.pdf` file that you downloaded previously.
3. Review the structure of the document.

   Notice that the document consists primarily of the following structures:

   - Title
   - Table of contents
   - Subtitles
   - Text

- Footnotes
- Bibliography

4. The SDU tool has predefined labels for all but the `footnotes` and `bibliography`. You will create new field labels for these two document structures in a later procedure.

Processing is finished when the page shows that one document is available.

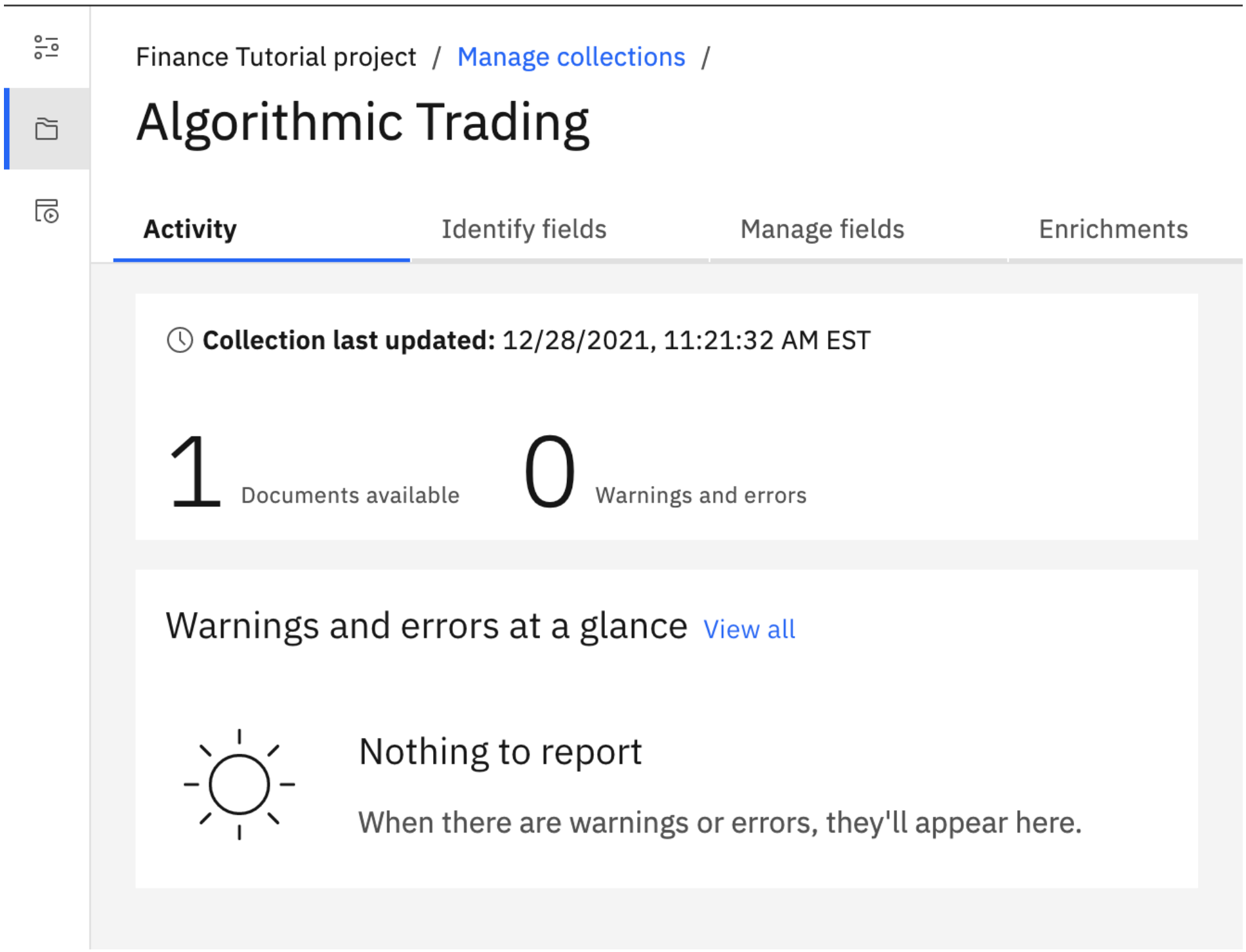

Figure 6. Activity page that shows the data upload is finished

Step 4: Test your project

1. After the crawl is completed, go to the *Improve and customize* page. From the navigation panel, click **Improve and customize**.
2. In the *Search* field, enter `When did the Flash Crash occur and why?`

   The following passage is returned as the response:

   ```
   These could in turn generate systemic destabilizing market events, such as the May 2010 “Flash Crash.” The “Flash Crash” occurred on May 6, 2010, when an algorithm rapidly sold 75,000 S&P500 e-mini futures contracts.
   ```

   The returned passage contains an accurate answer to the question.

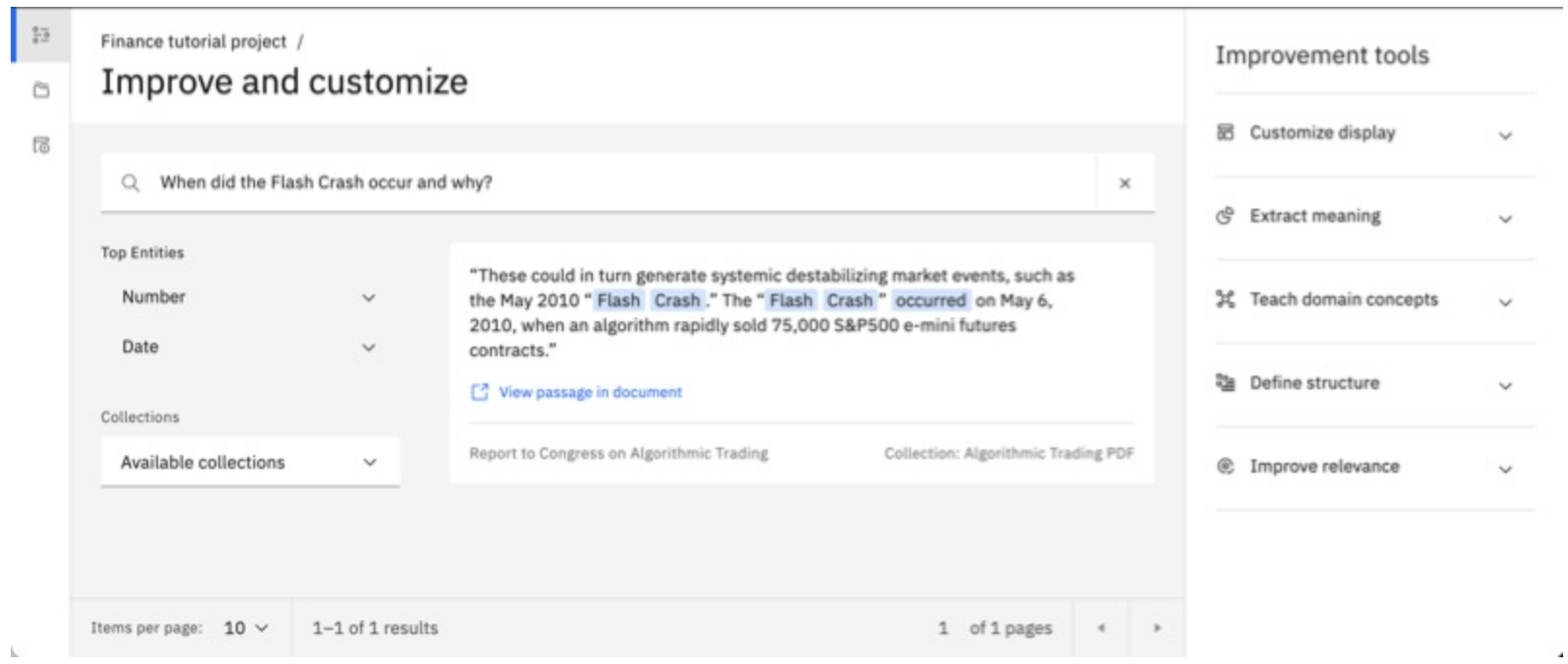

Figure 7. Search results

3. Ask another question, `What is the purpose of Rule 15c3-5?`

   The following passage is returned as the response:

   ```
   mechanism.306 b. 15c3-5 In November 2011, the SEC implemented the final provision of Rule 15c3-5 curbing
   unfiltered market access. The provision mandated that brokers verify their clients' order flow for compliance
   with credit and capital thresholds before routing to market centers
   ```

   Again, the answer is accurate (despite there being some extraneous text at the beginning of the passage).

   In both examples, a somewhat complex question is asked and the passage that is returned provides a valid answer.

   However, not every question returns as clear an answer. Next, we try some queries that generate answers we might want to improve.

4. Enter `Where do muni bond trades get reported to?`

   In this case, the response does not answer the question completely.

   ```
   Post-trade transparency, in the form of transaction reports, generally is available for corporate and
   municipal bonds. 1. Transaction Reports in Corporate Bonds: TRACE Transactions in corporate bonds must be
   reported to the Trade Reporting
   ```

5. Similarly, the search query, `What are PTFs?`, does not return a direct answer.

   ```
   Despite the surge in trading volume during the event window, there was no noticeable change in net positions
   of PTFs or bank-dealers. However, the report also finds evidence that some PTFs and bank-dealers may have
   contributed to the volatility
   ```

Your project is answering some of the questions successfully. Only one passage is being returned for each query. Let's see whether we can improve the responses that are given to these simpler search queries.

Step 5: Create a user-trained Smart Document Understanding (SDU) model

To improve the quality of the search results, build a Smart Document Understanding model for this document. The model helps Discovery understand the document structure. You can then instruct Discovery about which sections of the document to search and which sections to ignore.

1. From the *Improvement tools* panel of the *Improve and customize* page, expand *Define structure*, and then click **New fields**.

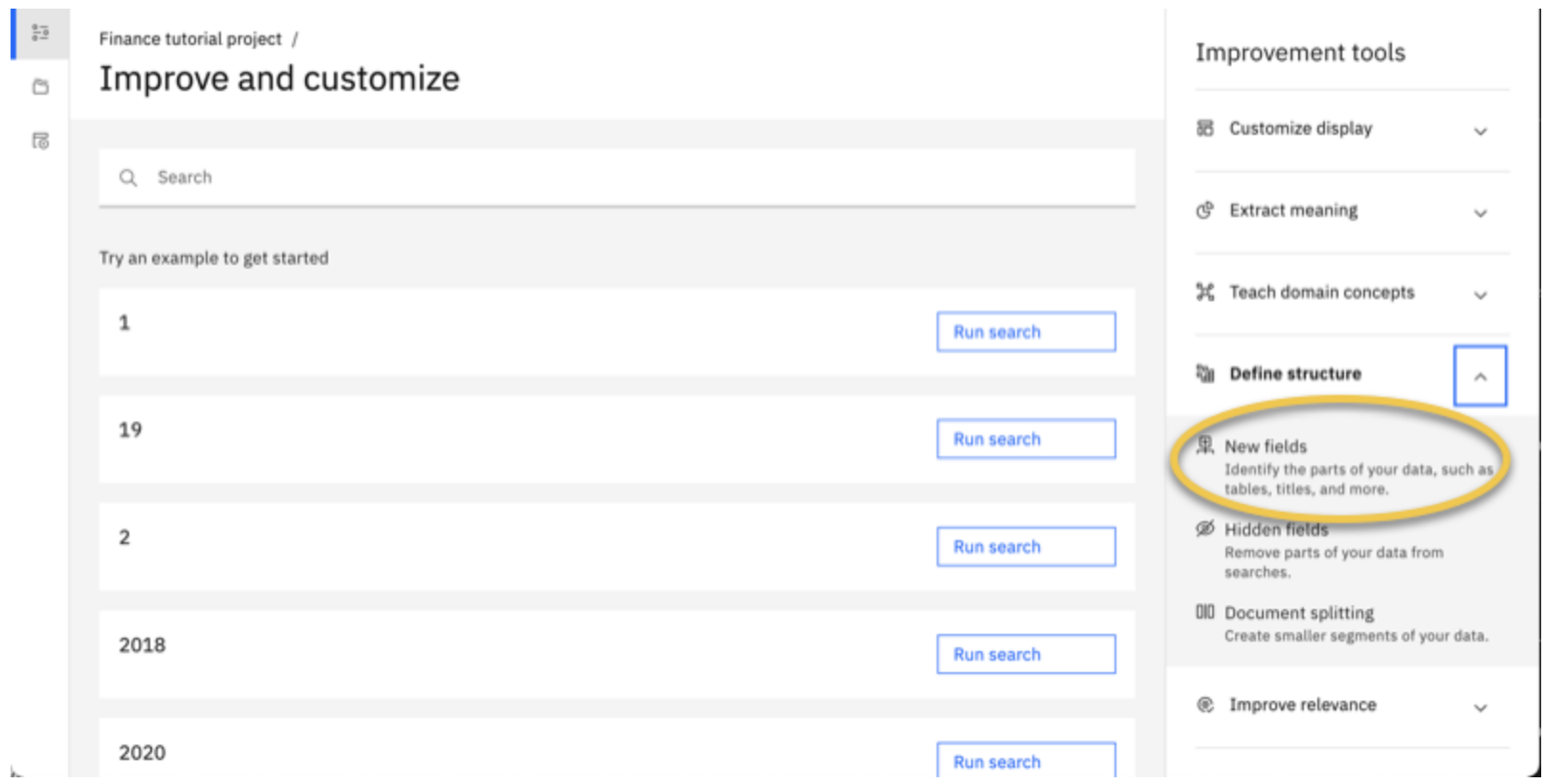

Figure 8. New fields tool in the Improvement tools panel

2. The *Identify fields* tab is displayed, where you can choose the type of Smart Document Understanding model that you want to use.

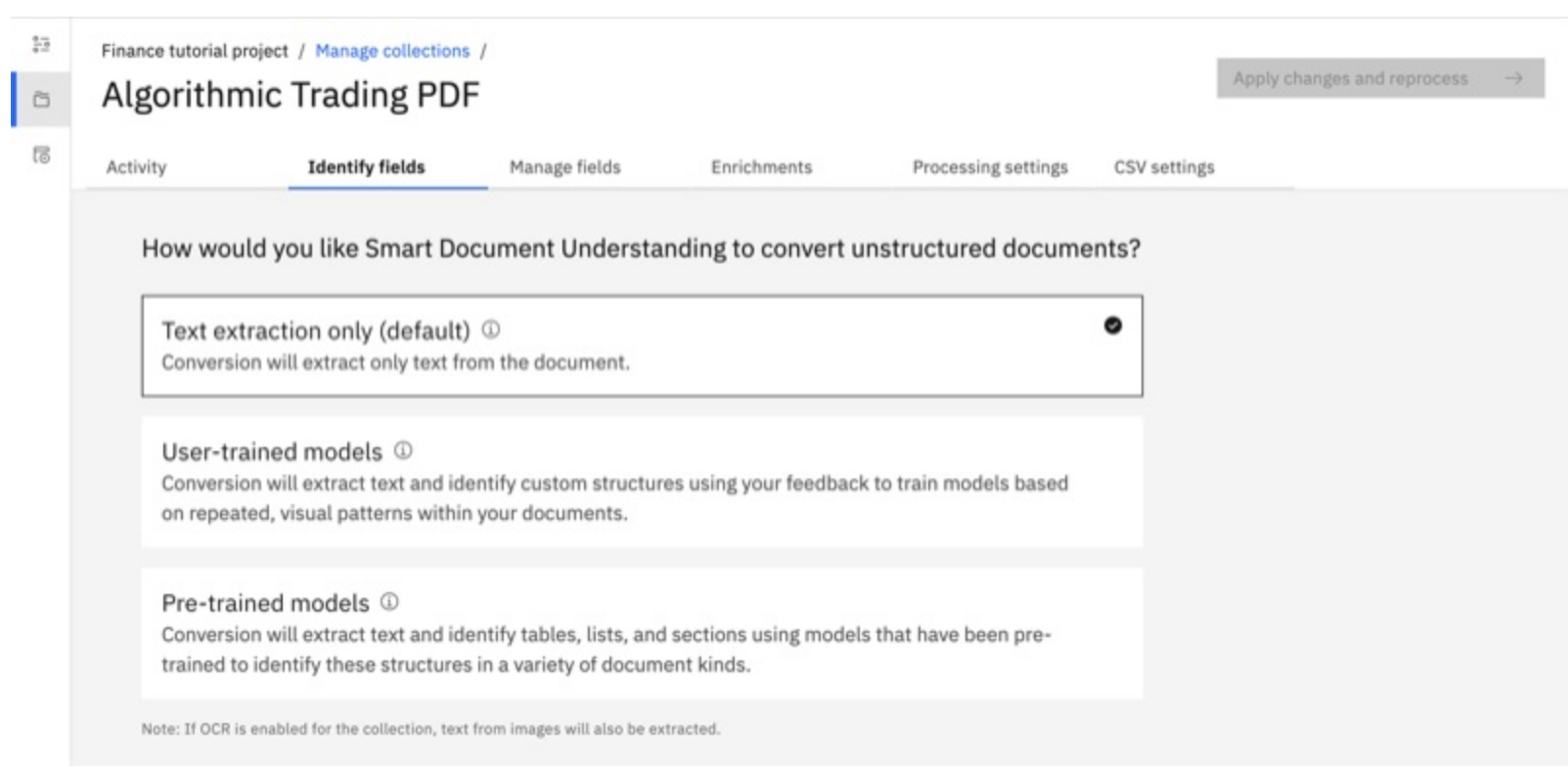

Figure 9. Identify fields tab

- The *pretrained model* applies a noncustomizable model that extracts text and identifies tables, lists, and sections. The pretrained model is a great choice to save time.
- For the purposes of this tutorial, where we want to explore how the Smart Document Understanding tool works, we'll choose to use the *user-trained model*.

If you don't choose a model, the *text extraction* model is applied automatically. With the text extraction model, most of the document content is treated as standard text and is indexed in the `text` field.

3. Click **User-trained models**, and then click **Submit**.

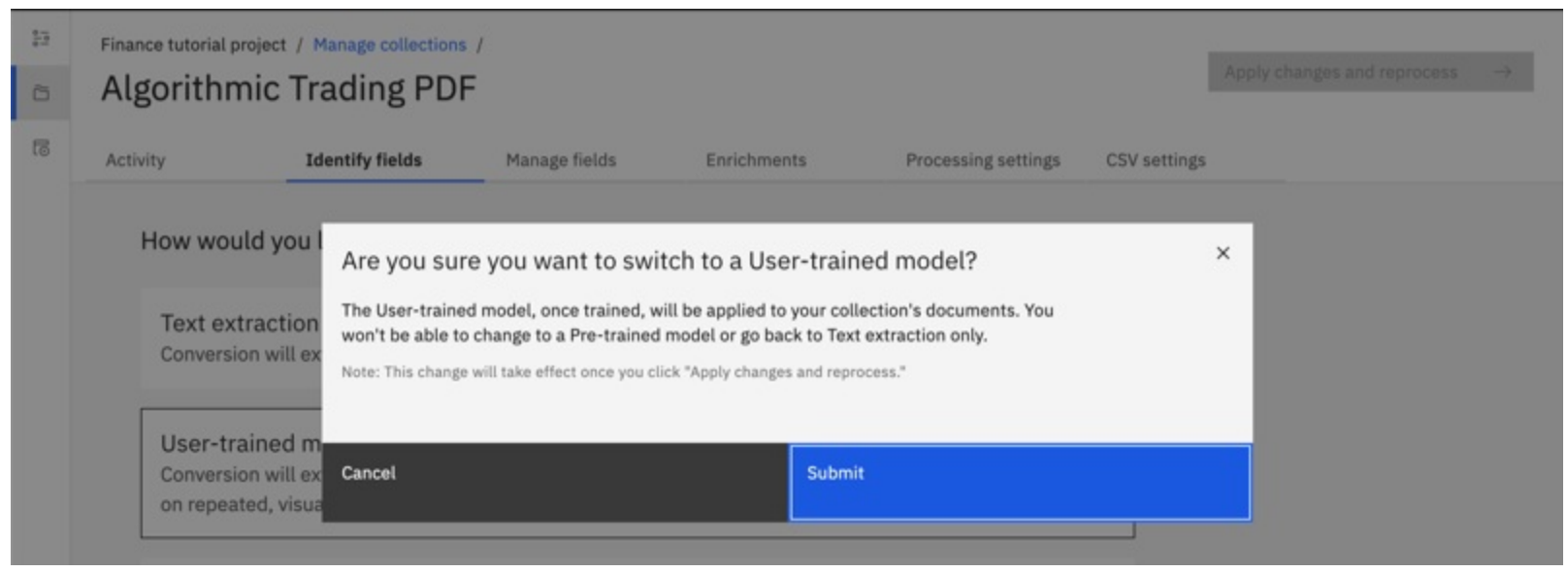

Figure 10. Confirmation dialog for user-trained model

4. Click **Apply changes and reprocess**.

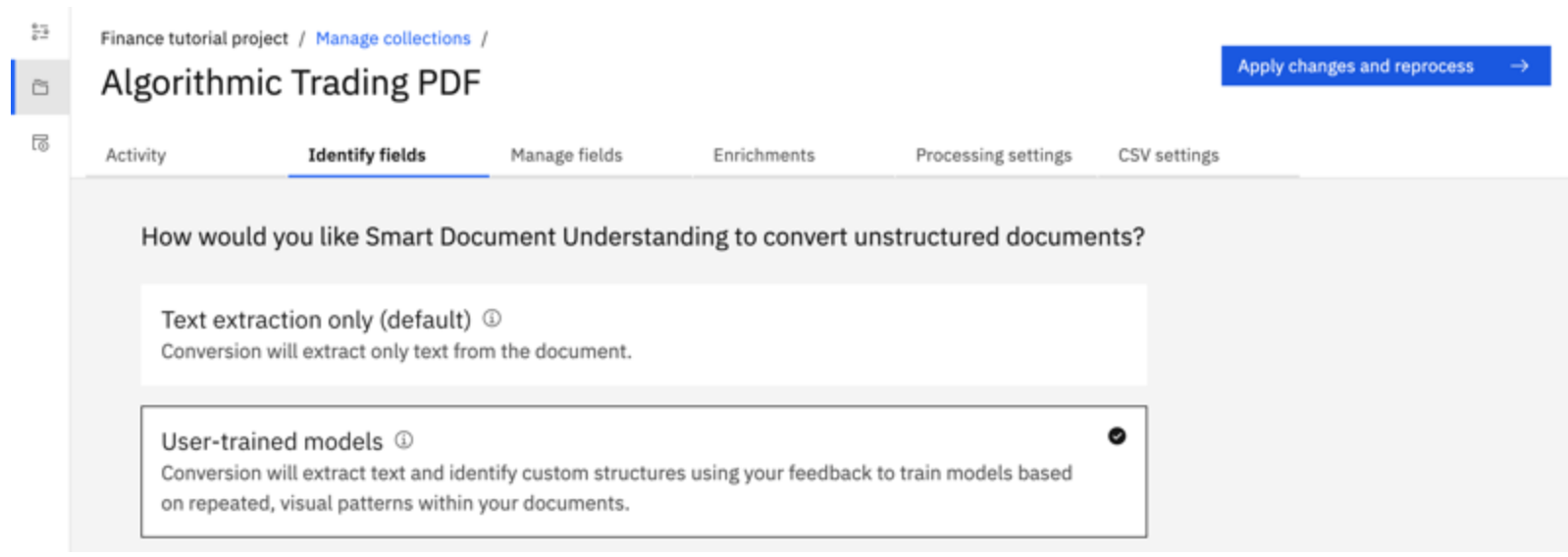

Figure 11. Apply changes and reprocess button

After the evaluation process is complete, a representation of the document is displayed in the Smart Document Understanding tool.

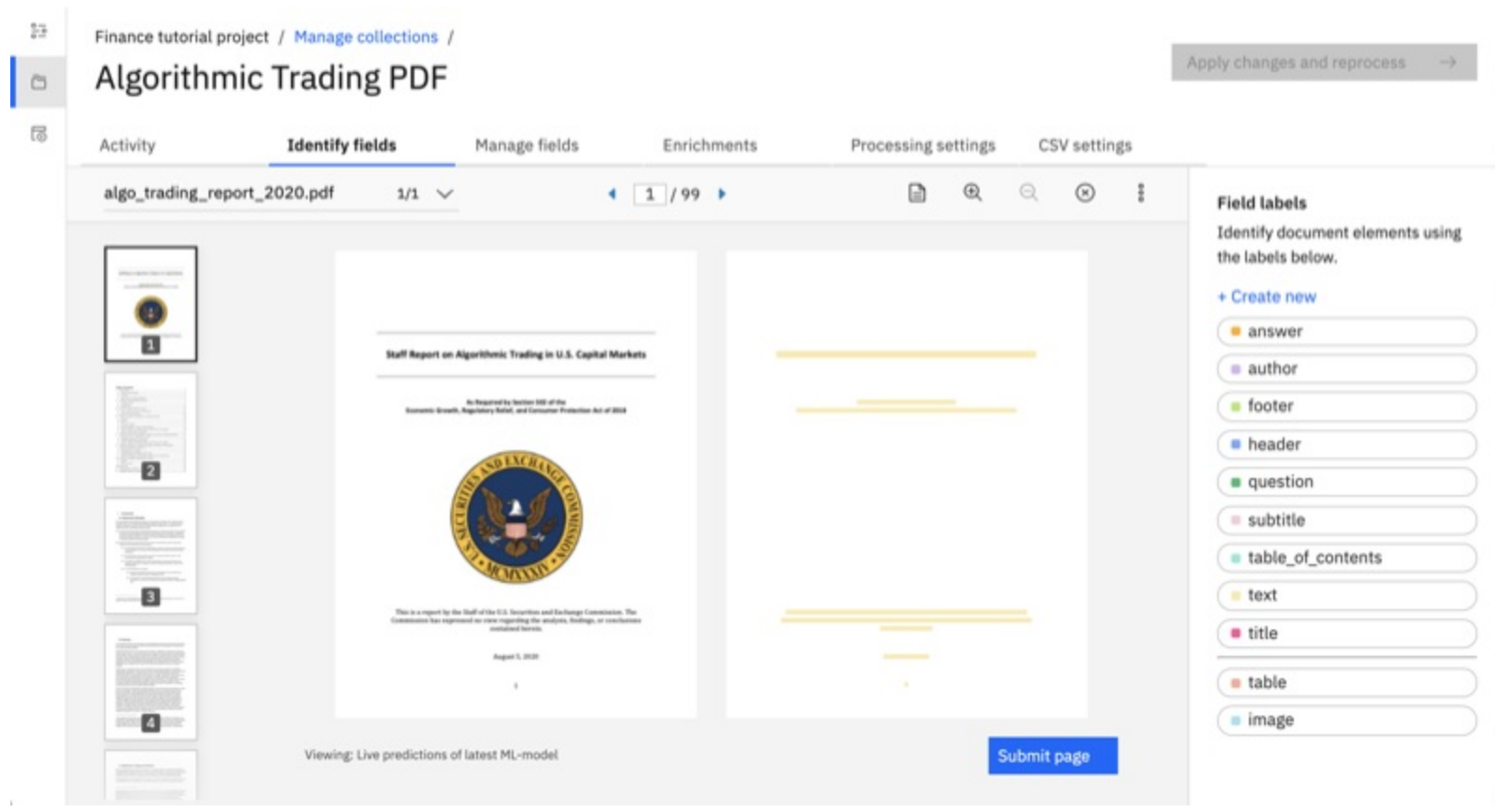

Figure 12. PDF is displayed in the SDU tool

The tool shows you a view of the original document along with a representation of the document, where the text is replaced by blocks. The blocks represent field types.

Initially, the blocks are all the color of the `text` field label because all of the document content is considered to be standard text and will be indexed in the `text` field.

A *Field labels* list shows the predefined field labels that are available.

We are going to label blocks that represent specific types of information, such as titles and subtitles, with corresponding field labels. (The process of using labels to identify different parts of the document's structure is called *annotating* the document.)

5. To annotate the document, click the label first. Then, click the block of text that you want to label.

   Click `title` from the *Field labels* list, and then, in the document representation, click the yellow block that is situated in the location of the document title.

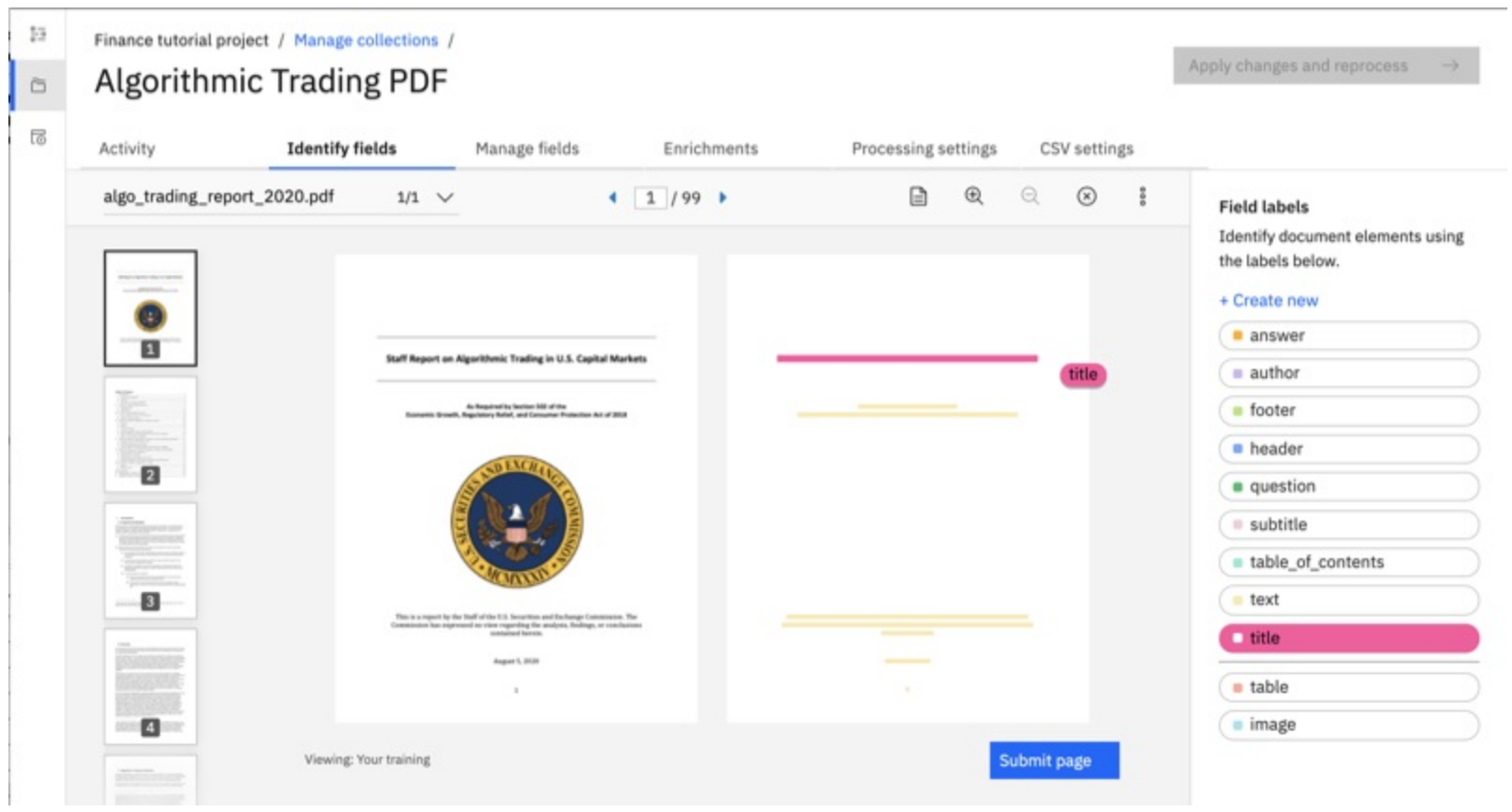

Figure 13. A title is being labeled in the Smart Document Understanding tool

You labeled the title of the document successfully!

6. The rest of the text on the page can be indexed as part of the `text` field. Therefore, click **Submit page**.
7. The next page is the *Table of contents* for the document. Click the `table_of_contents` label, and then select all of the text on the page to label it. (You can click and drag the mouse to select all.) Click **Submit page** to move to the next page.

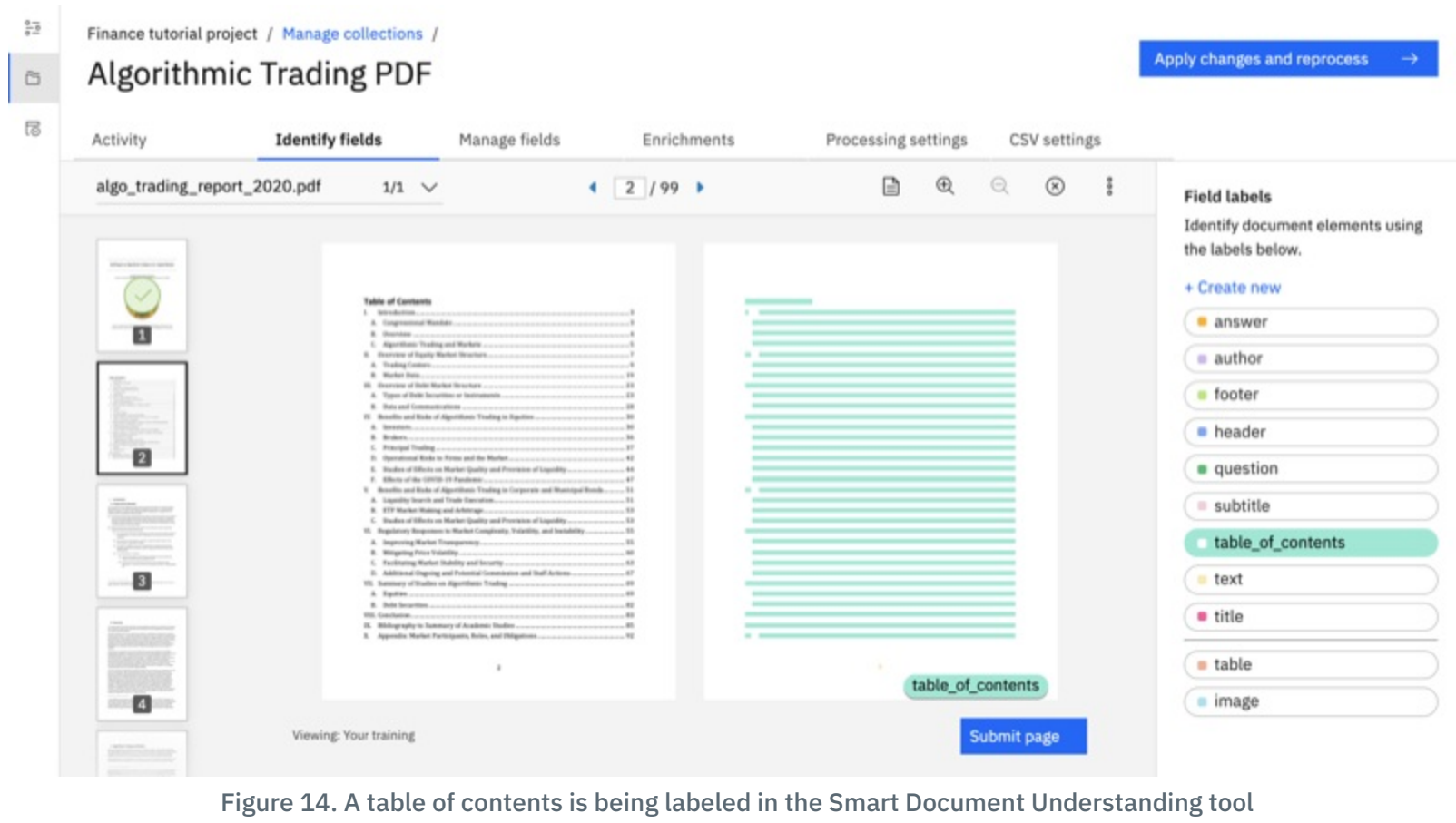

Figure 14. A table of contents is being labeled in the Smart Document Understanding tool

8. The two headings on the page are subtitles. Click the `subtitle` label, and then select the headings.

   This page has a footnote. As we noted earlier, the document has many footnotes where some important information is provided. Let's label the footnotes so we can include or exclude this type of information later. There is no footnote label, so we must add one.

9. From the *Field labels* list, click **Create new**. Add the name `footnote` as the label name. Click the color block repeatedly until you find a unique color to use for the label, and then click **Create**.

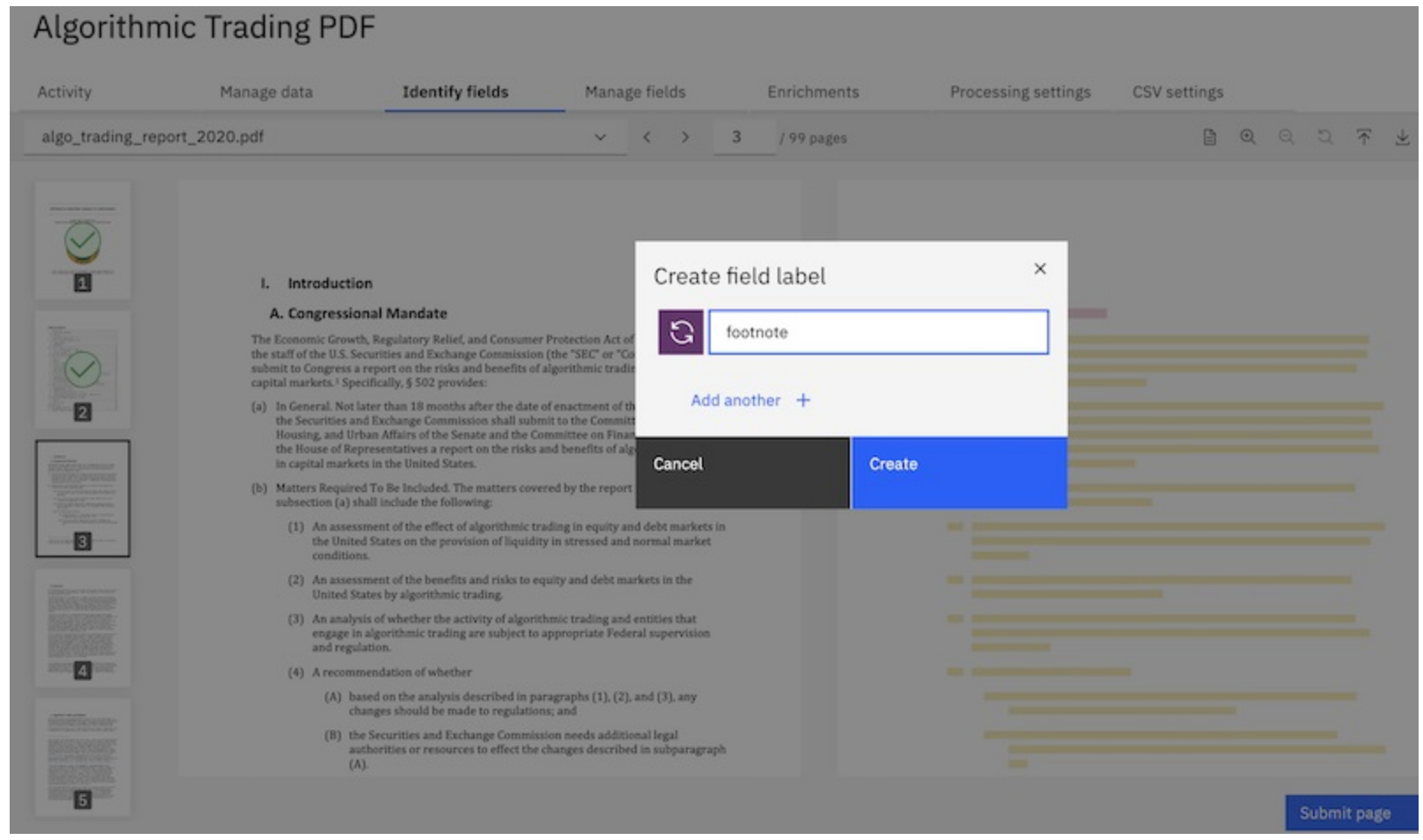

Figure 15. New label dialog

10. Click the new footnote label that you added, and then label the footnote on the page with the label. Click **Submit page** to move to the next page.

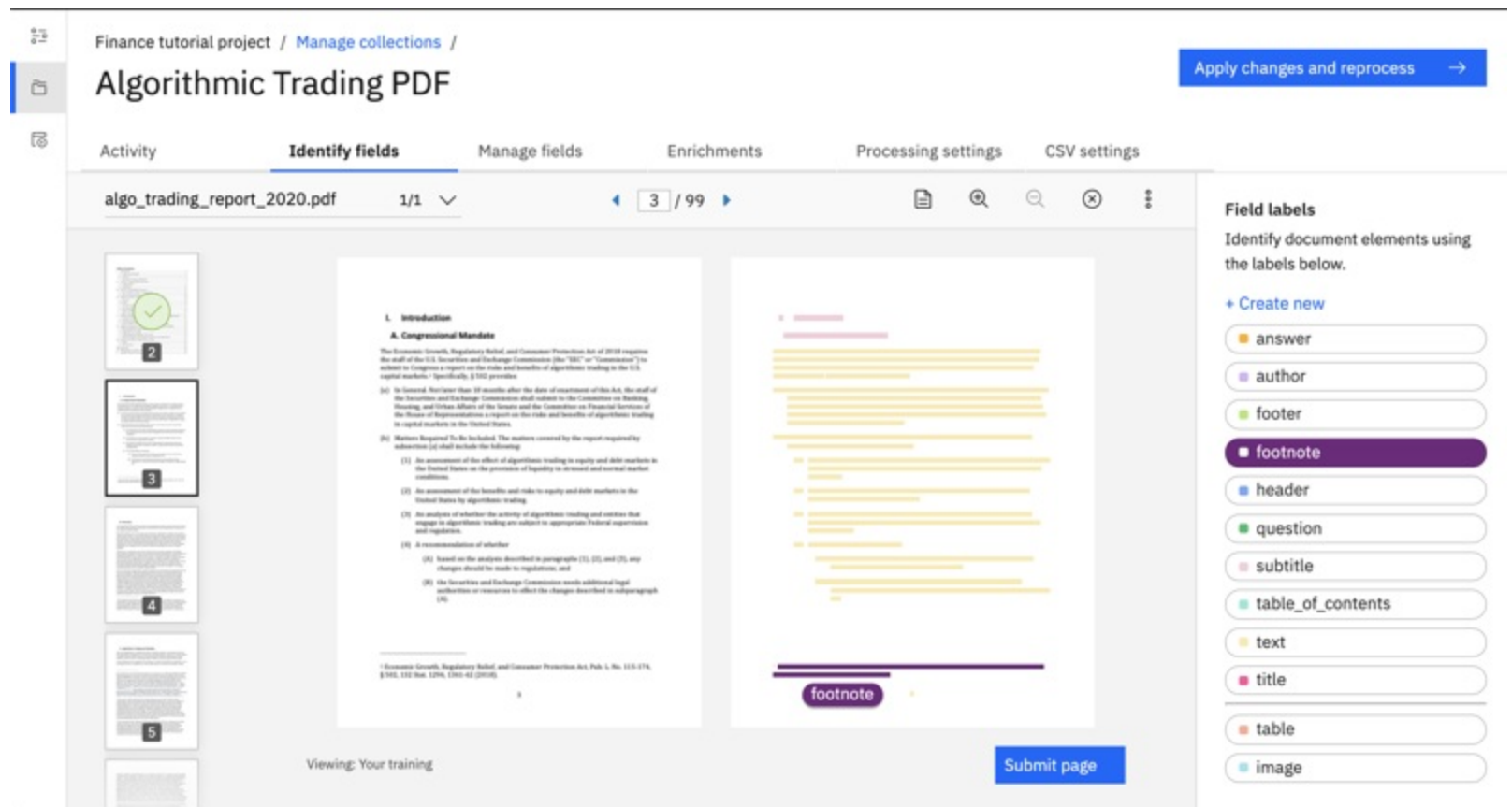

Figure 16. A footnote is being labeled in the Smart Document Understanding tool

11. Repeat this process to label and submit multiple pages.

    For most pages, the content includes a `subtitle`, a `footnote`, and the bulk of the content on the page remains labeled as `text`.

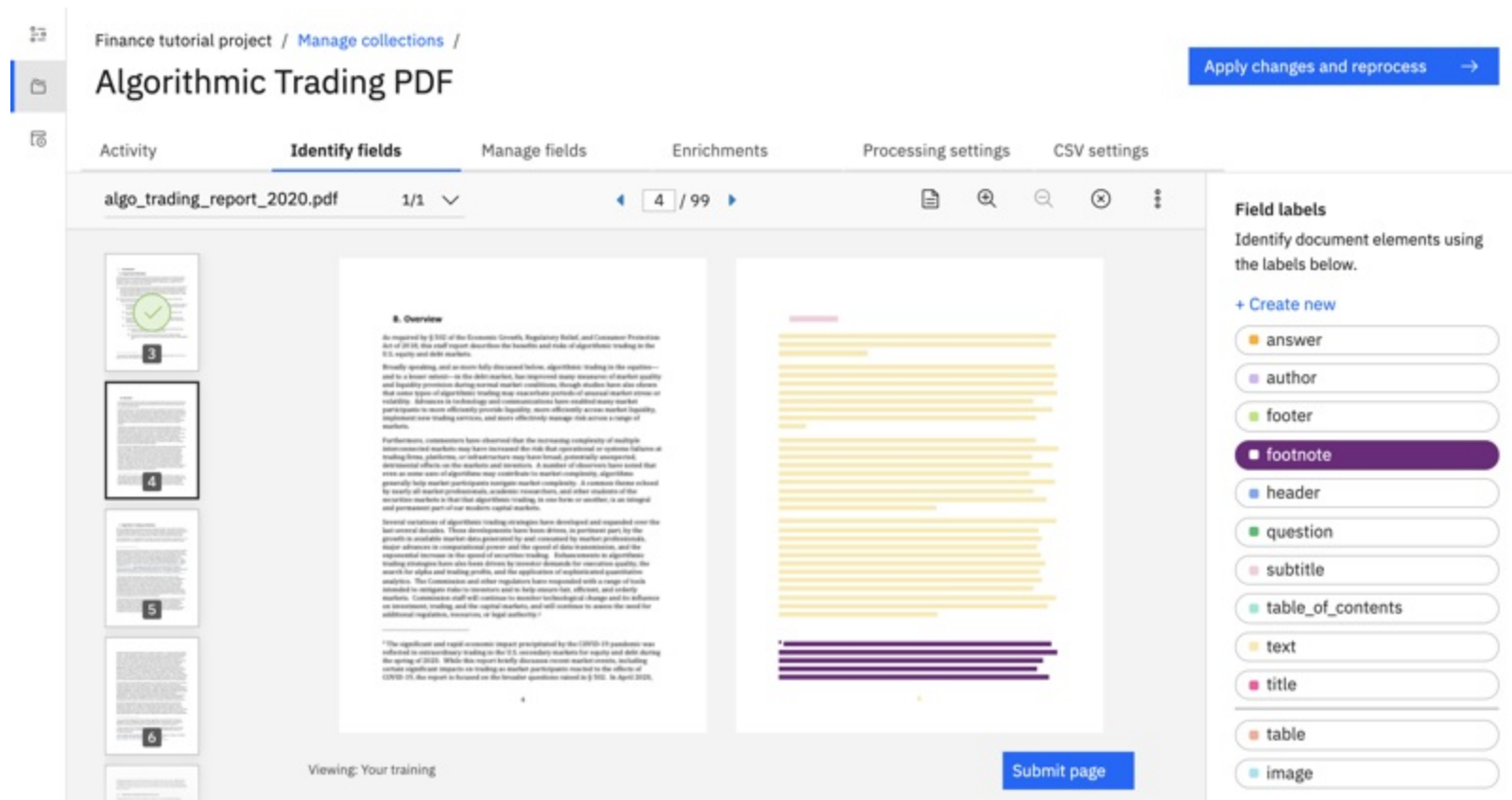

Figure 17. Subtitle, footnote, and text labels are being applied

As you label and submit pages, the model learns from the annotations that you make. Gradually, the labels that are applied automatically become accurate and don't require any adjustments.

If the tool becomes overzealous in the application of labels, apply the `text` label to sections of standard text to correct it.

12. For tables, select the table caption and the entire table and label them with the `table` label.

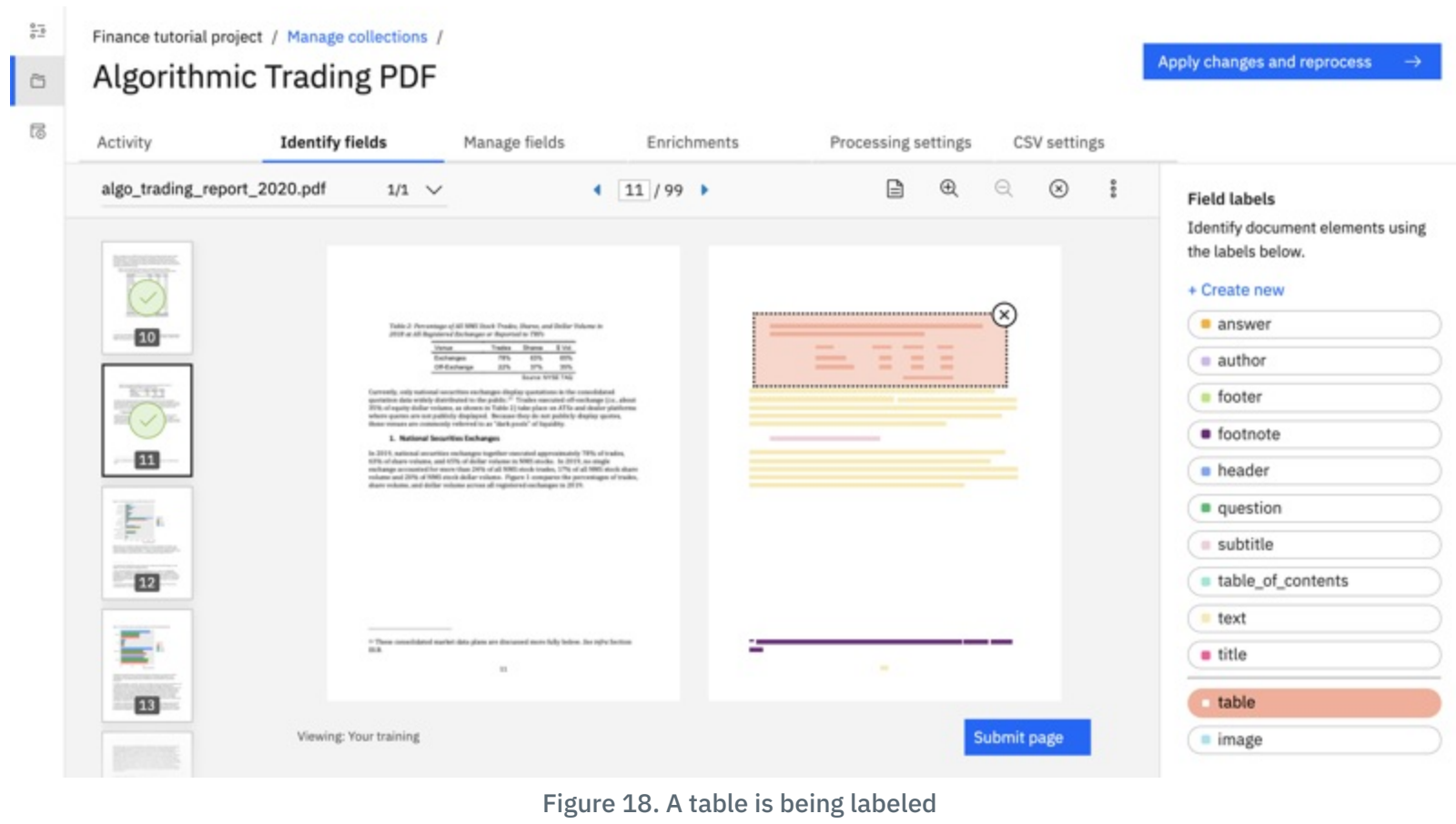

Figure 18. A table is being labeled

13. When a page contains an image, the image is not displayed in the representation of the page.

    Images are never replicated. However, you can capture the text from an image so that the image text can be searched. To do so, enable the Optical Character Recognition (OCR) feature when you create a collection. OCR is helpful in cases where you want to extract text from images, such as from a scanned PDF, where the text is embedded in an image. For more information, see Optical character recognition.

    After you enable OCR, if you want to remove annotated image text from the collection index, you can label the image so that you can exclude the associated text. You will learn about how to configure the index in the next procedure.

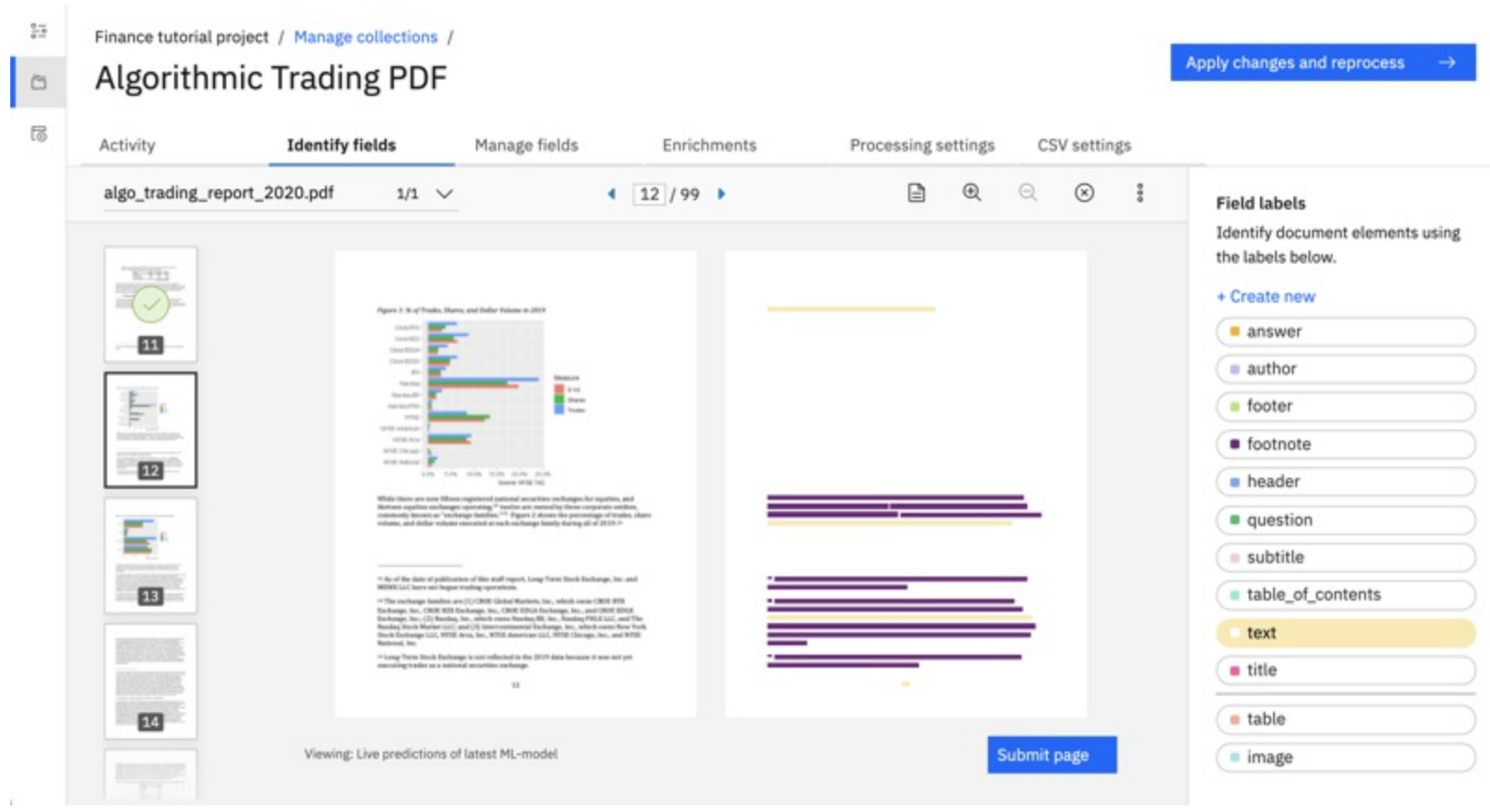

Figure 19. Shows an image in the page

14. When you reach the *Bibliography*, create a new label named `bibliography`.

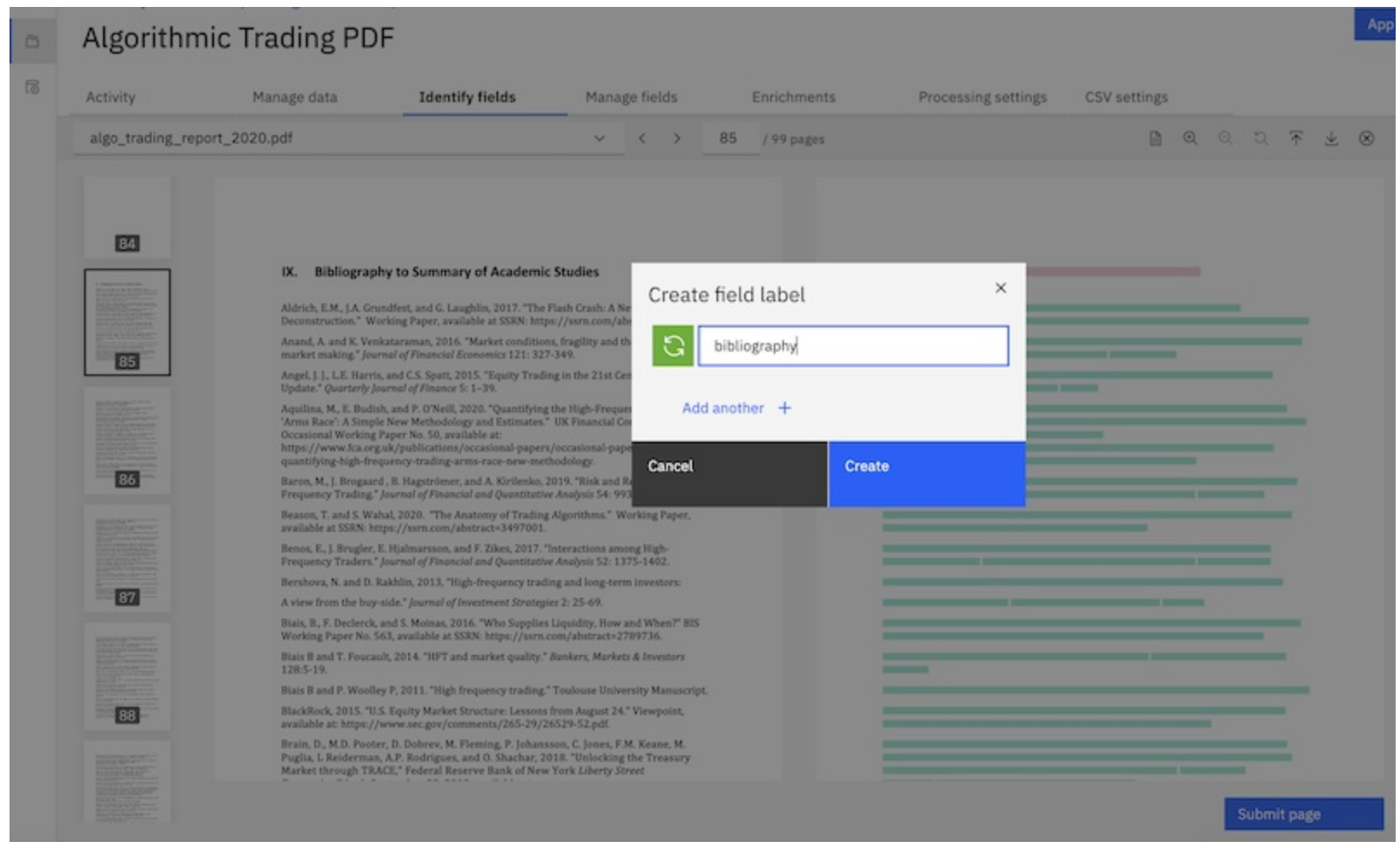

Figure 20. Creating a bibliography label

Apply the new label to each page.

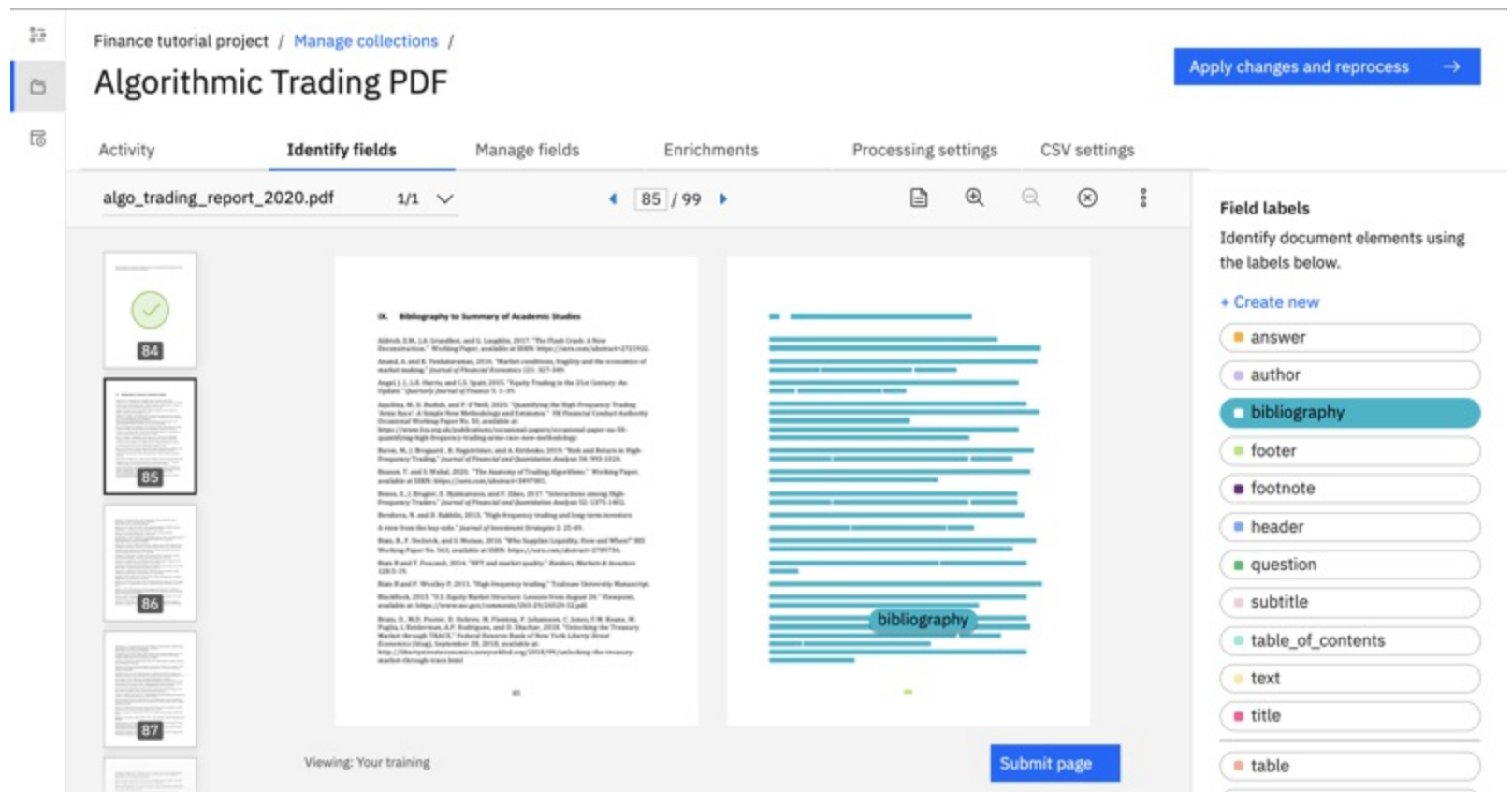

Figure 21. A bibliography label is being applied

15. After you annotate and submit all the pages, click **Apply changes and reprocess**.

    A notification is displayed to indicate that the collection was updated. You remain on the SDU tool page, but the *Apply changes and reprocess* button is disabled.

An SDU model is generated based on the structures that you labeled in this document.

For more information about the Smart Document Understanding feature, see Using Smart Document Understanding.

Step 6: Streamline the searchable data

Now that you have an SDU model that can recognize the different types of sections in the document, you can instruct it to include some sections in searches and to exclude others. To control what data gets searched, you include or exclude fields from the search index.

1. Click **Manage fields**.

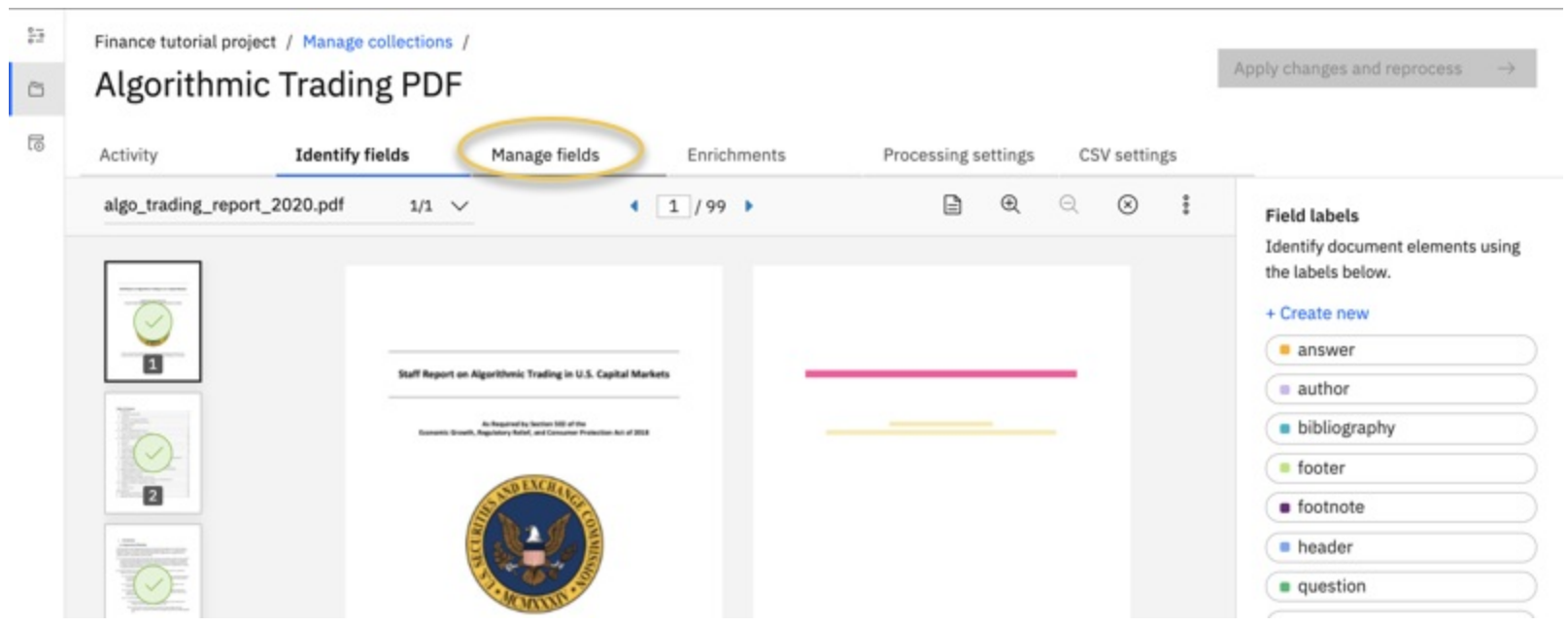

Figure 22. The Manage fields tab

2. From the list of fields to index, set the switcher to **No** for all fields except these ones:
    - `footnote`
    - `html`
    - `subtitle`
    - `table`
    - `text`

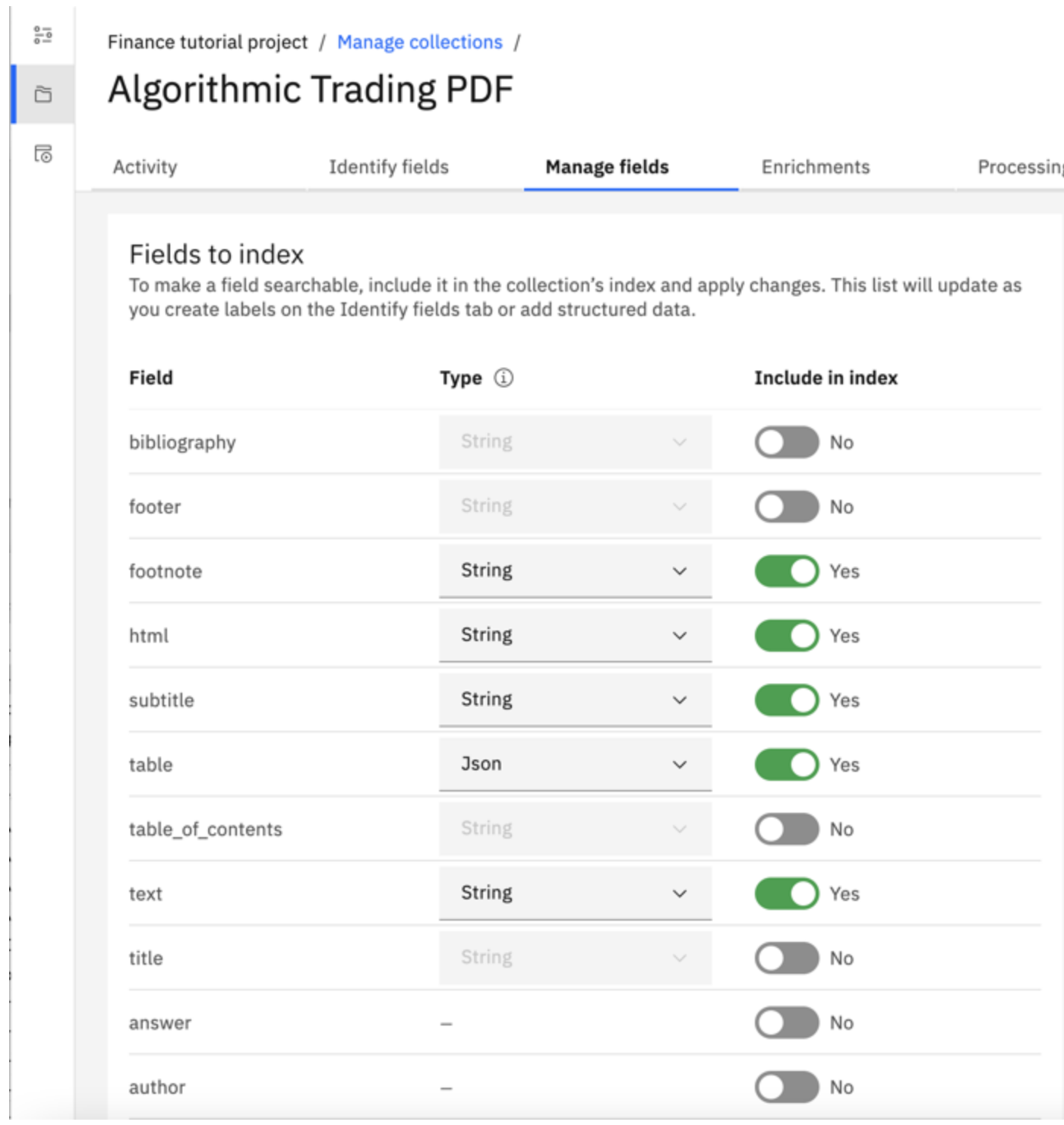

Figure 23. Fields in the index list

3. Click **Apply changes and reprocess**.

   A notification is displayed to indicate that the collection was updated. You remain on the *Manage fields* page, but the *Apply changes and reprocess* button is disabled.

You successfully configured the index to control the content that is available to searches! You excluded fields that might contain popular search terms, but do not also include meaningful content.

For more information about managing fields, see Excluding content from query results.

Step 7: Split the document

Now that Discovery knows more about the structure of the document, we can split the single 99-page document into more documents. Remember, only one passage was returned for each query that you submitted before. If we split the document into multiple segments, Discovery can return the best passages from across all of the document segments.

> **Note:** When you split a document, you turn one document into many documents. Be aware of the document limits for your plan type. Each document segment that is generated by splitting a document counts toward the plan's document limit.

When you annotated the document, you identified the `subtitle` field. These subtitles are a good marker from which each new document segment can begin.

1. From the *Improve query results by splitting your documents* section of the *Manage fields* page, click **Split document**.
2. Select `subtitle` from the *Split document on each occurrence of* field.

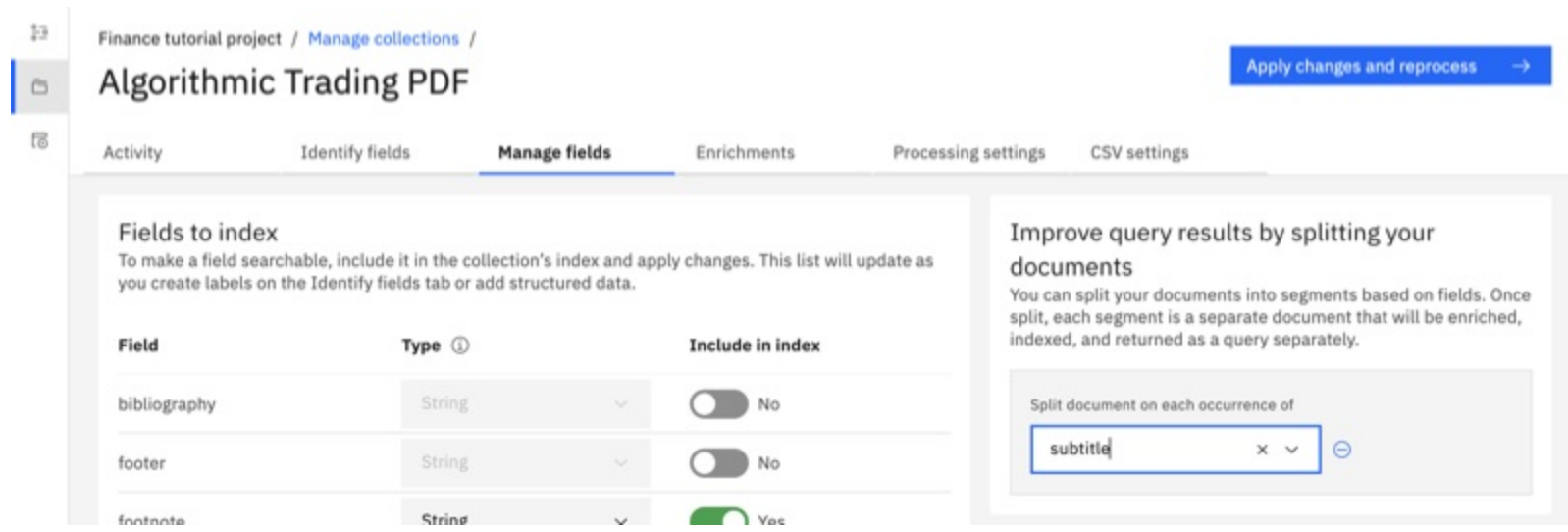

Figure 24. Choosing to split documents on the subtitle field

3. Click **Apply changes and reprocess**.

   A notification is displayed to indicate that the collection was updated. You remain on the *Manage fields* page, but the *Apply changes and reprocess* button is disabled.

4. Click **Activity** from the page header to return to the *Activity* page where you can monitor the progress of the change you made.

When no documents are processing, document splitting is finished.

For more information about splitting documents, see Split documents to make query results more succinct.

Step 8: Test the project again

Let's find out whether we improved the search function by adding a user-trained SDU model for the document. To do so, let's retest the project.

1. From the navigation panel, click **Improve and customize** to open the *Improve and customize* page.
2. First, to make sure that we didn't degrade the quality of the search, let's ask one of the questions that returned a good response when we tested earlier.

   In the *Search* field, enter `What is the purpose of Rule 15c3-5?`

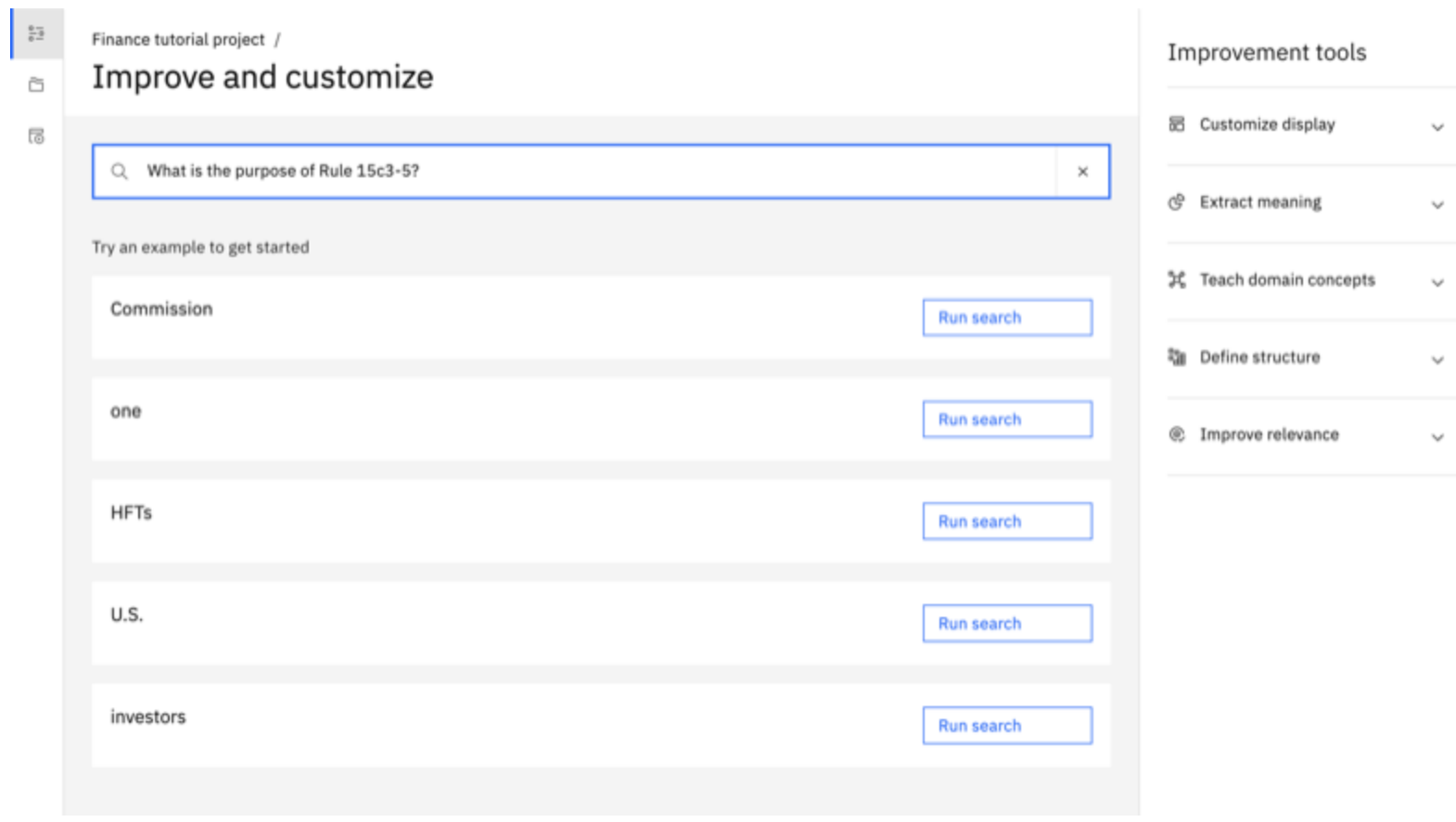

Figure 25. Query added to the Improve and customize page

Multiple responses are returned this time. The following response contains the exact answer to the question without any extraneous text:

```
In November 2011, the SEC implemented the final provision of Rule 15c3-5 curbing unfiltered market access. The provision mandated that brokers verify their clients' order flow for compliance with credit and capital thresholds before routing to market centers.
```

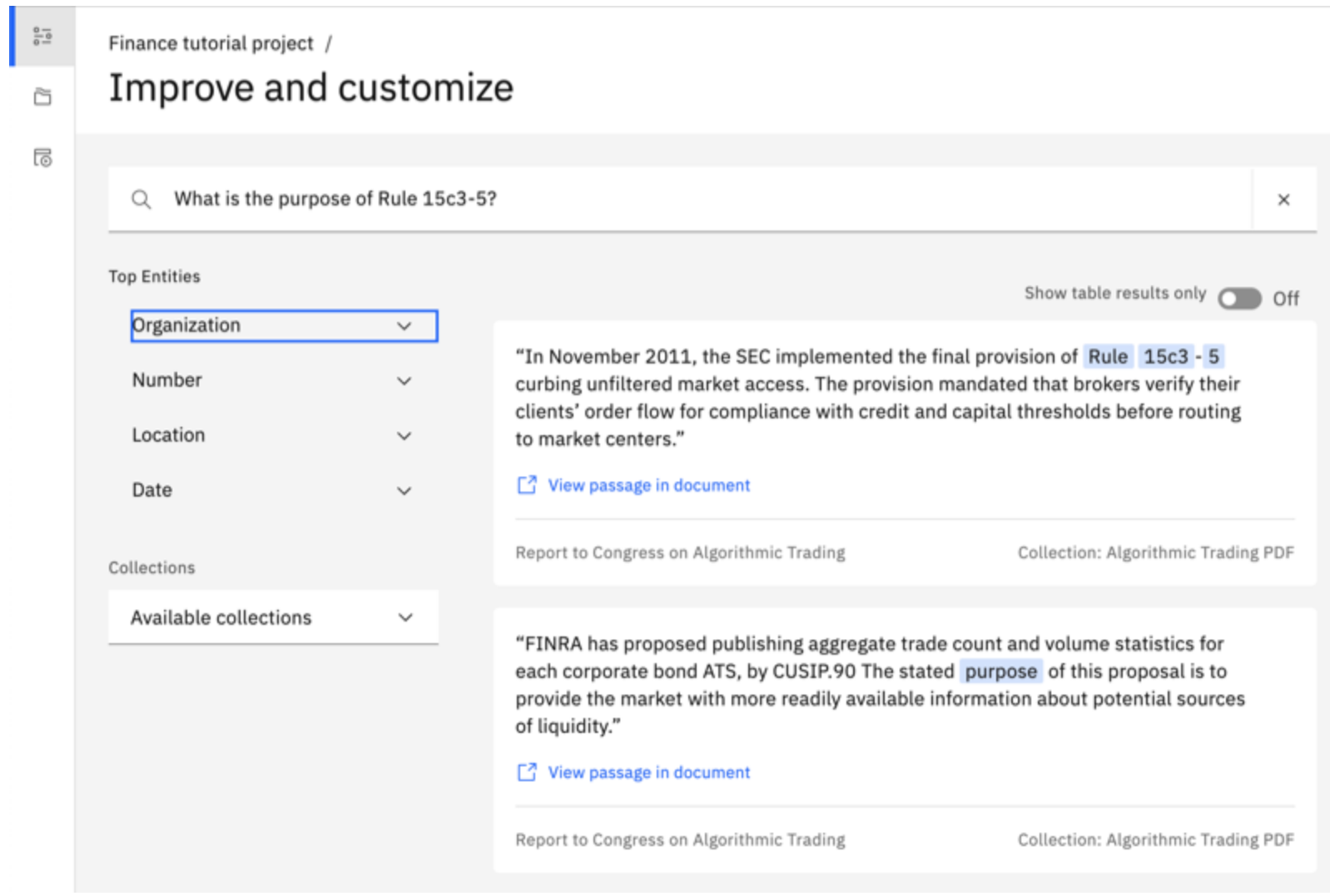

Figure 26. Multiple responses are returned for the query

Our updates only improved the quality of the accurate responses that were returned before.

3. Now, let's ask a question that returned poor results previously. Enter `What are PTFs?` as the search query.

   The same response that was returned as the only response last time is returned again. However, this time we get more than one response. And we can see that the second response that is returned defines the acronym for us.

   `("principal trading firms" or "PTFs")`

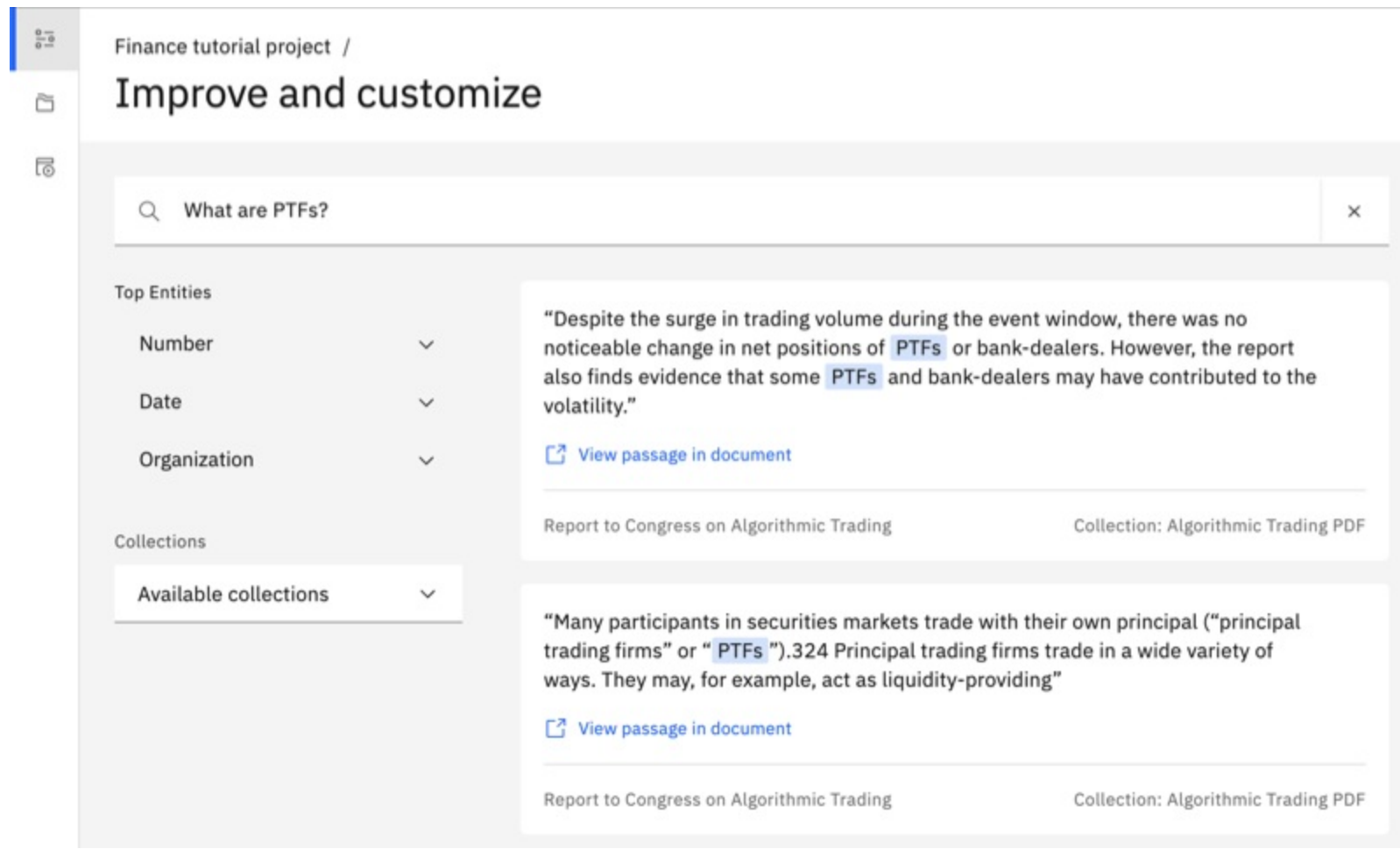

Figure 27. Responses that answer the question about PTFs

4. Let's try the other problematic search query. Enter `Where do muni bond trades get reported to?` as the search query.

   This time it's the third response that provides an answer to the question. You must view the full passage to see the entire definition.

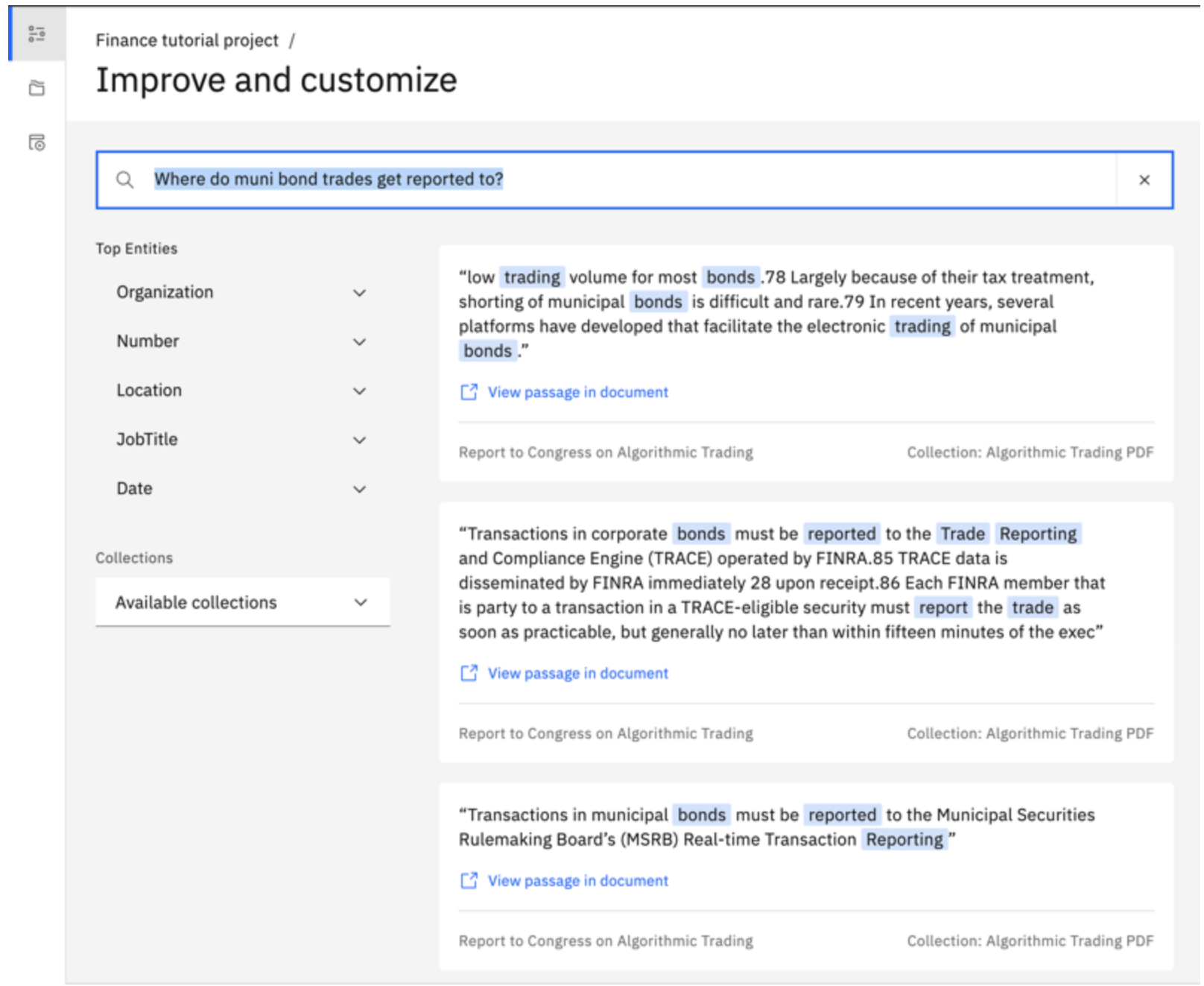

Figure 28. Responses that answer the question about muni bonds

Click the *View passage in document* link to see the full definition highlighted in the document.

```
Transactions in municipal bonds must be reported to the Municipal Securities Rulemaking Board's (MSRB) Real-time Transaction Reporting System (RTRS).
```

Congratulations! You successfully added a user-trained Smart Document Understanding (SDU) model that improves the quality of your search project.

Step 9: Filter results with a dictionary-based facet

Now that we are getting more passages returned per query, it might be useful to filter the results. To filter the results based on the types of financial instruments that are mentioned, we can add a search facet. One available source for a facet is a dictionary.

1. To create a dictionary, from the *Improvement tools* panel of the *Improve and customize* page, expand *Teach domain concepts*, and then click **Dictionaries**.

2. Click **New**.

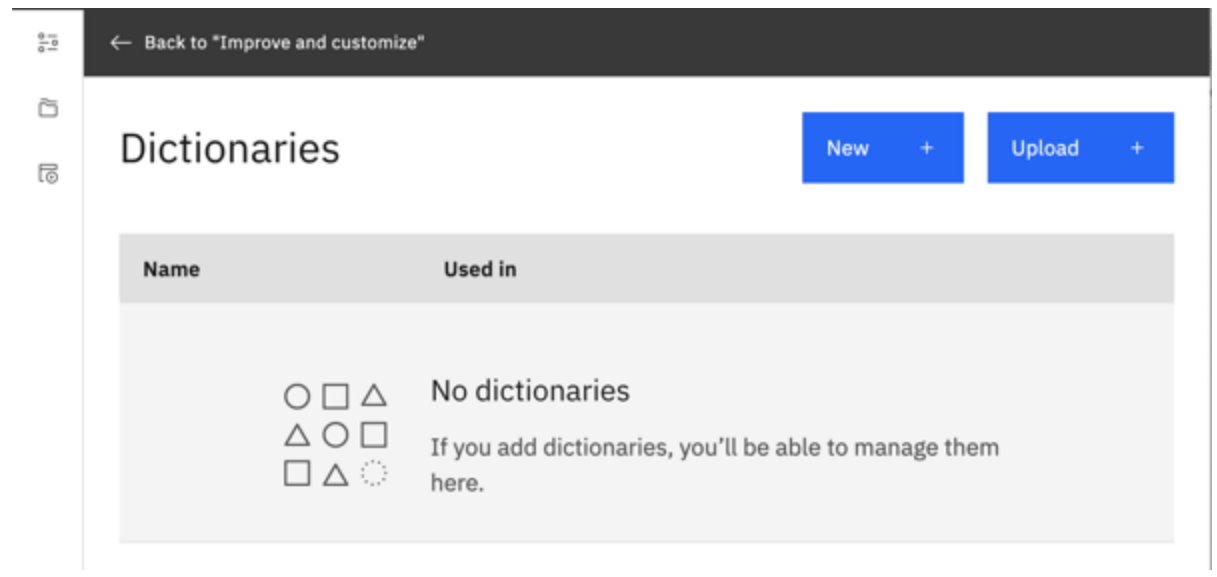

Figure 29. New button in the dictionary page

3. Enter `Financial instruments` as the dictionary name, add the term `municipal bond`, and then click the *Add term* button.

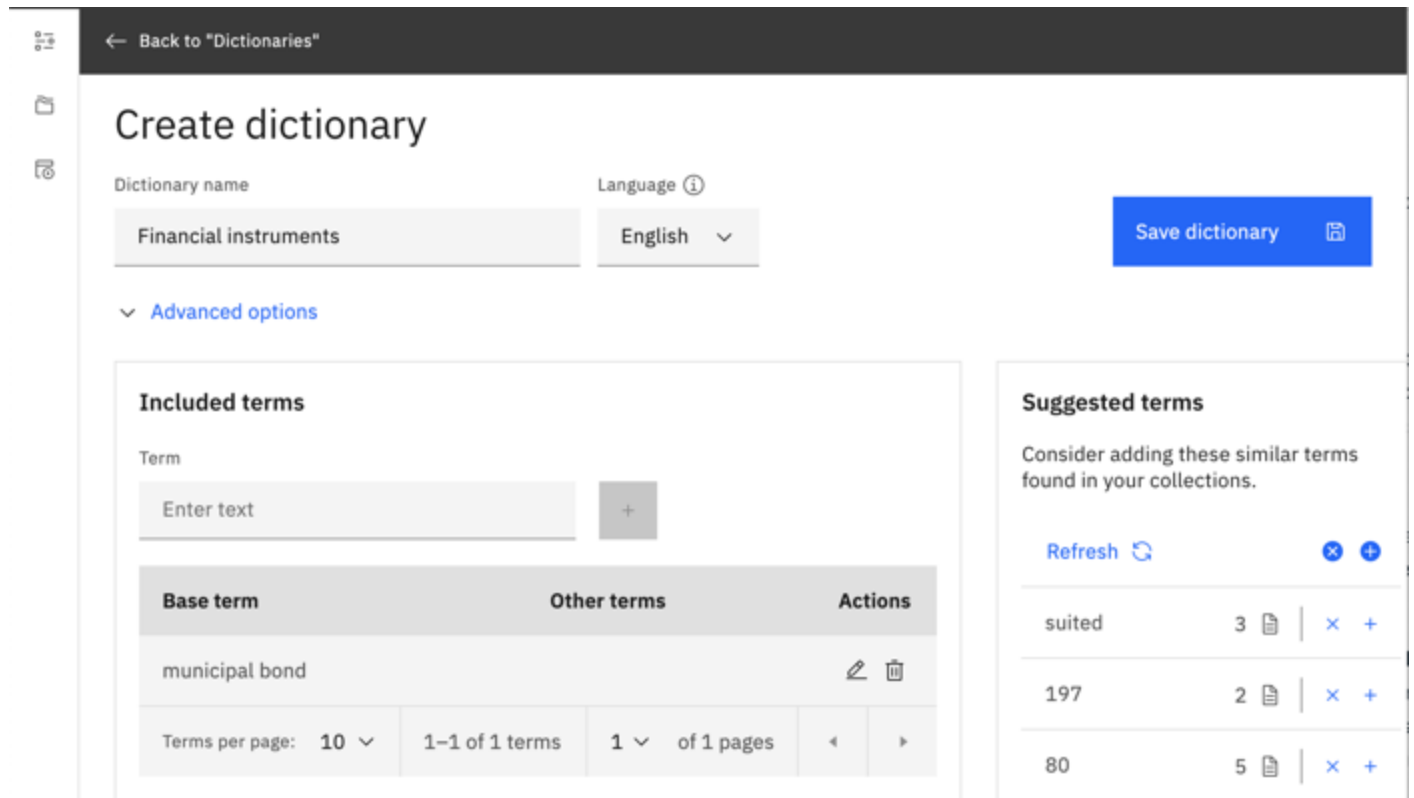

Figure 30. Financial instruments dictionary with one term

4. Add synonyms for the term by clicking the edit icon for the term.

   ```
   Municipal Bonds, muni, munis, muni bonds
   ```

   Add synonyms in a comma-separated list, and then click **Save term**.

5. Click **Save dictionary**.

   You can choose a field in the document where you want the enrichment to be applied. Let's choose the `subtitle` field that was generated when we created the user-trained SDU model. From the *Fields to enrich* field, select `subtitle`. Click **Apply**.

   The dictionary is created and each subtitle in the document is analyzed for mentions of terms or synonyms that are defined in the dictionary. Any mentions that are found are noted in the index.

6. Click **Improve and customize** from the navigation panel.

7. From the *Improvement tools* panel of the *Improve and customize* page, expand *Customize display*, and then click **Facets**.

8. Click **New facet**, and then select **From existing fields in a collection**.

9. Choose the index field that is associated with the dictionary enrichment that you applied to the `subtitle` field. From the *Field* field, select `enriched_subsection.entities.mentions.text`

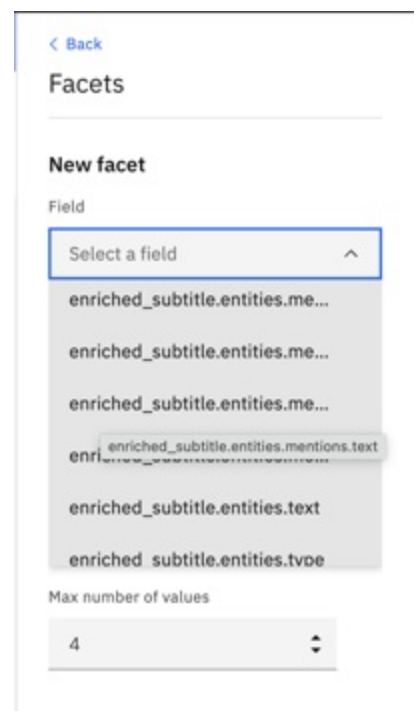

Figure 31. Fields from which you can create a facet

   You might need to hover over the entries to see the full field names.

10. Add a label, such as `Dictionary terms` to the *Label* field, and then click **Apply**.

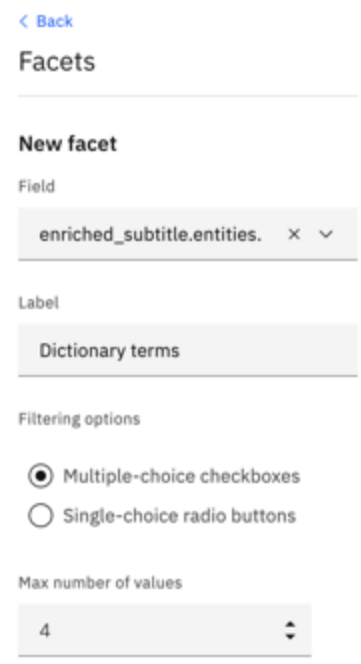

Figure 32. Facet was created

11. Enter `Where do muni bond trades get reported to?` as the search query.

    The *Dictionary terms* facet that you created is displayed along with the search results. A `Municipal Bonds` checkbox is shown, which indicates that at least one of the returned passages is extracted from a document segment with the term `Municipal Bonds` in its `subtitle` field.

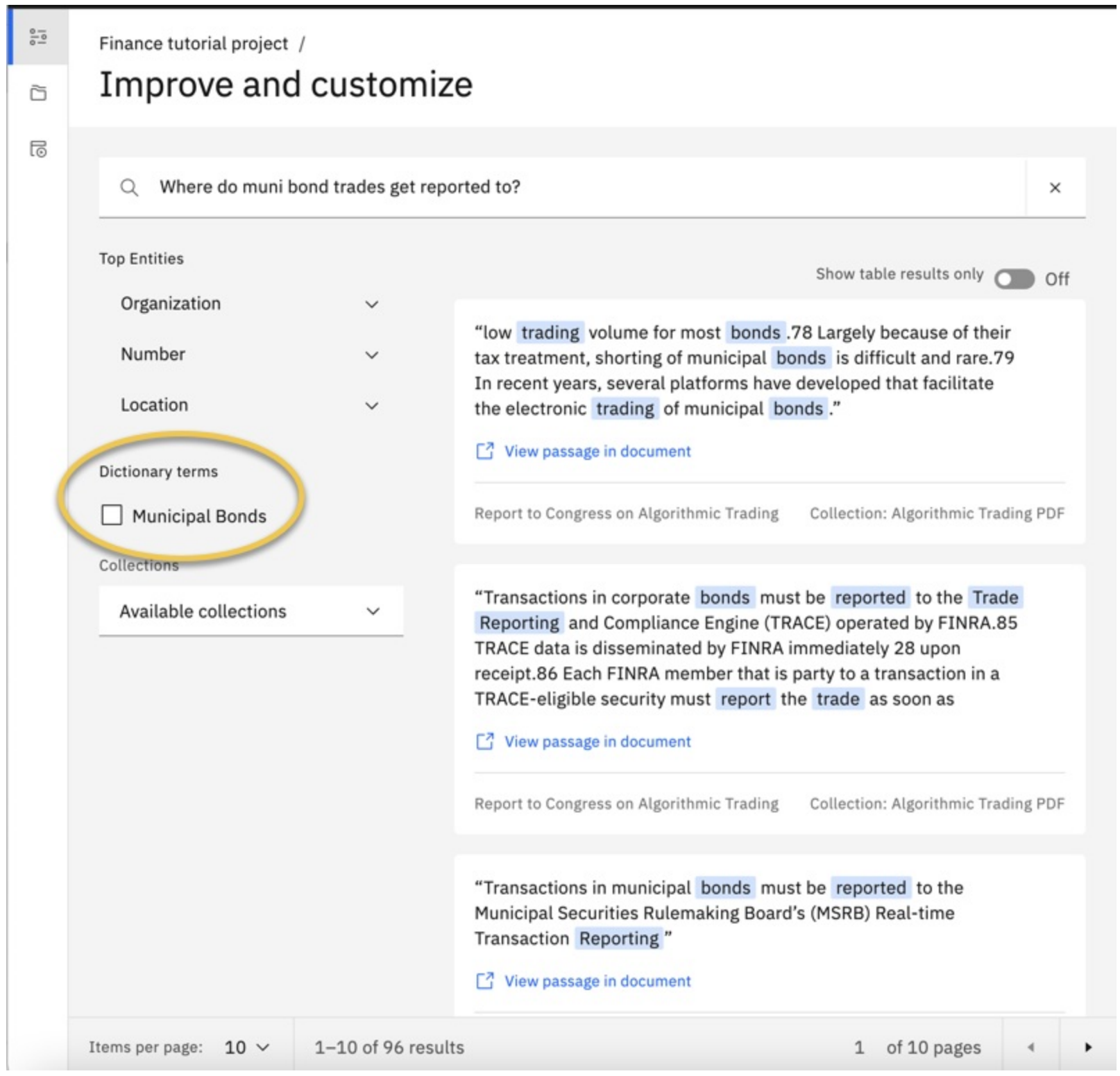

Figure 33. Dictionary term facet with a Municipal Bonds option

12. To filter the results to show only passages from sections with `Municipal Bonds` in the subtitle, select the `Municipal Bonds` checkbox.

    The best answer is now listed as the second response instead of the third.

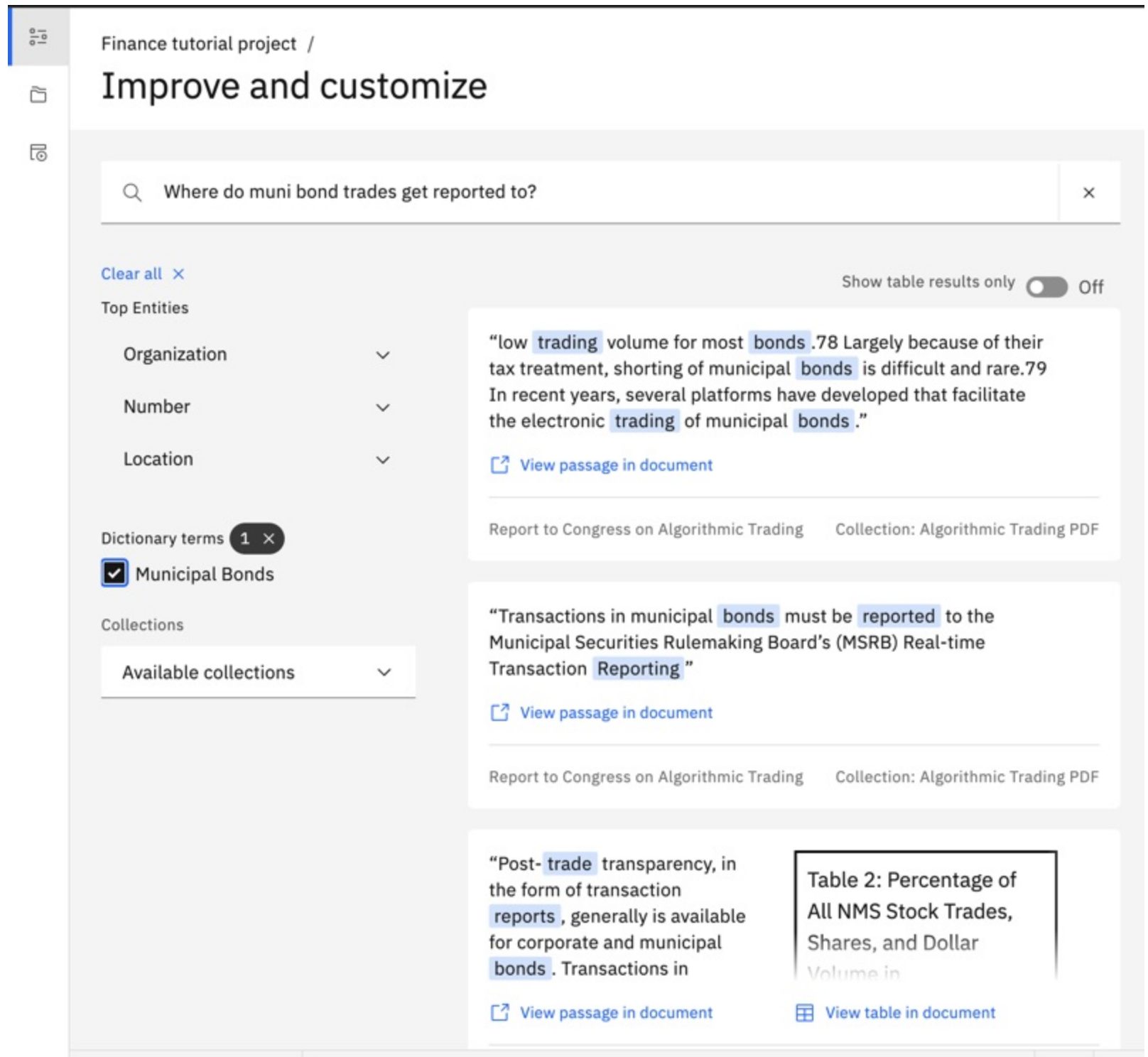

Figure 34. Best answer is the second result

Summary

In this tutorial, you created a Document Retrieval project, a Smart Document Understanding (SDU) model, a dictionary enrichment, and a search facet. You applied the facet that is based on your dictionary to the custom field that is generated by your SDU model to filter your query results for better answers.

Build an external webhook enrichment solution in Watson Discovery

In this tutorial, you can use sample applications to build an external webhook enrichment solution by using Watson Discovery.

IBM Cloud

> **Note:** Follow this tutorial only if you are using a managed deployment.

The following image shows the external enrichment configuration flow.

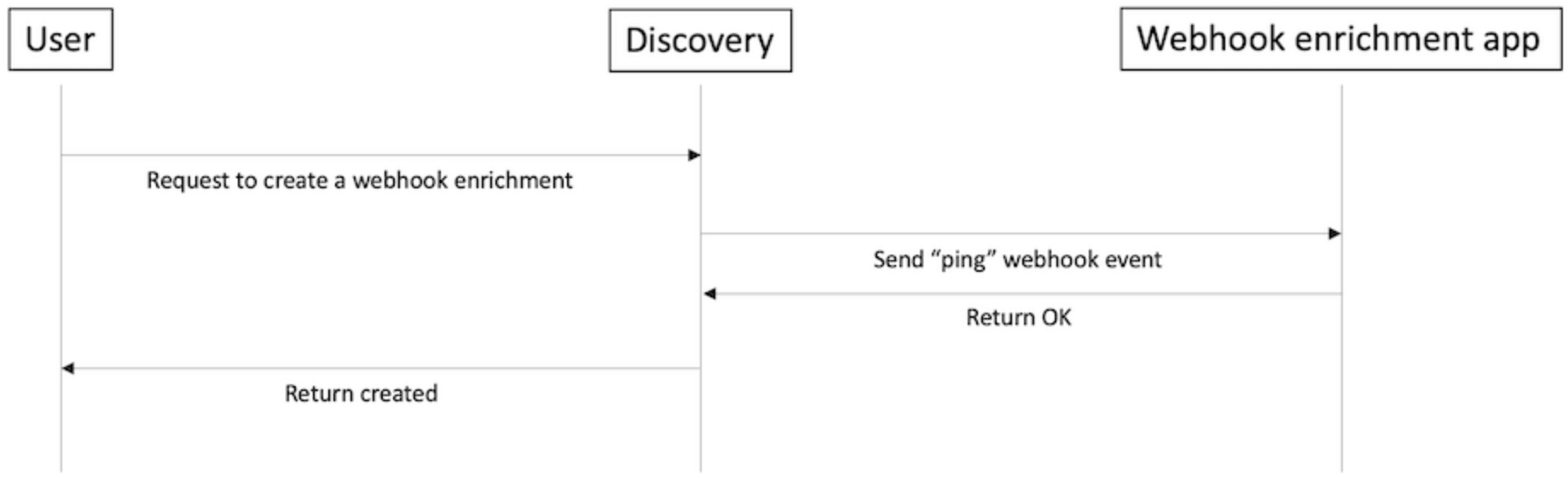

Figure 1. External enrichment configuration flow

The following image shows the external enrichment process flow.

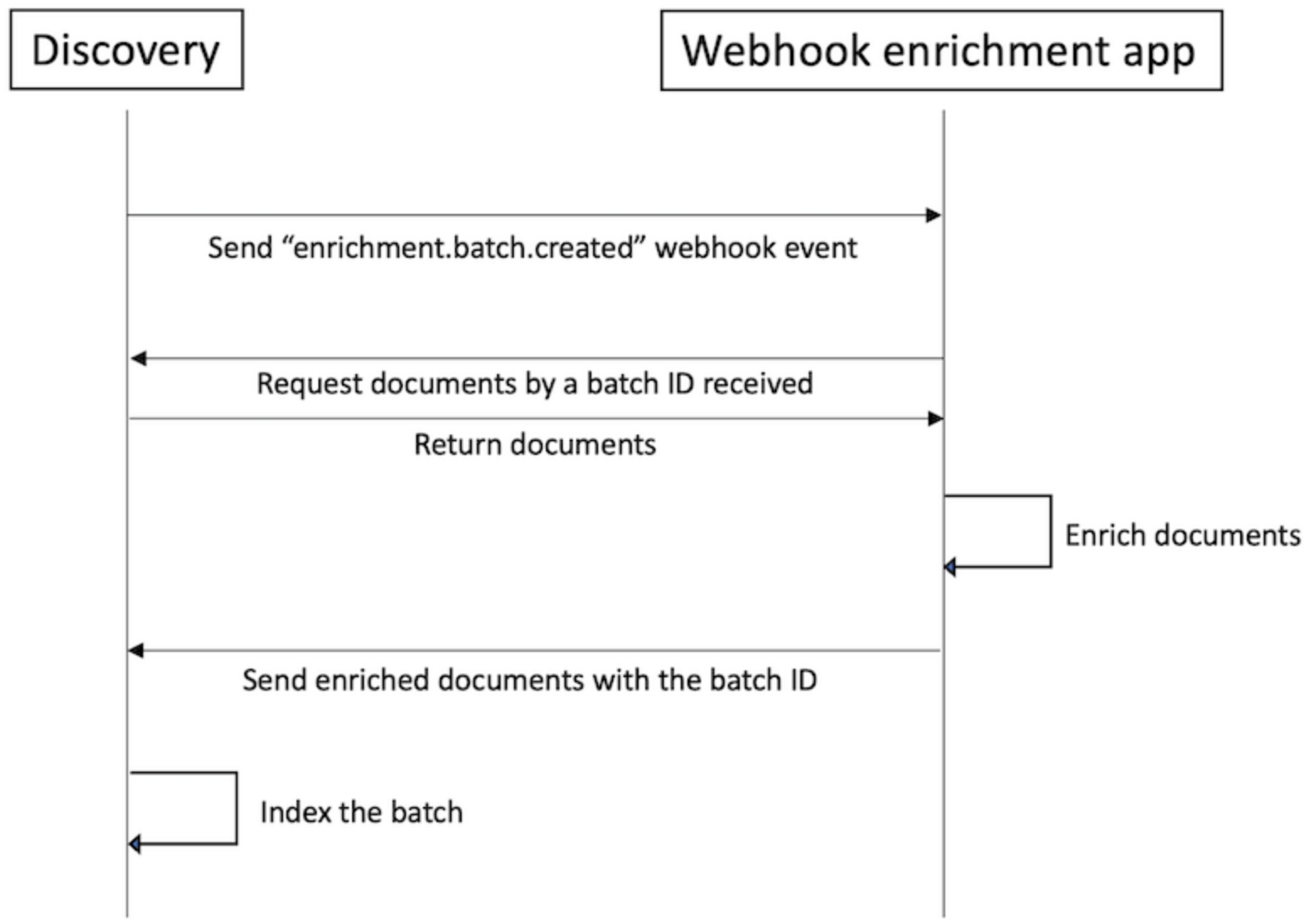

Figure 2. External enrichment process flow

For more information about the external enrichment APIs, see External enrichment API.

Learning objectives

By the time you finish the tutorial, you will learn how to use the following sample applications:

- `Regex`: For entity extraction, document classification, and sentence classification by using regular expressions
- `Granite`: For entity extraction from email by using watsonx.ai Granite model
- `Slate`: For entity extraction with watsonx.ai Slate model that is fine-tuned with labeled data exported from entity extractor workspace of Watson Discovery.

Duration

This tutorial takes approximately 2 hours to complete.

Prerequisite

1. Before you begin, you must set up a paid account with IBM Cloud to get an instance of Watson Discovery Plus or Enterprise plan.

   You can complete this tutorial at no cost by using a Plus plan, which offers a 30-day trial at no cost. However, to create a Plus plan instance of the service, you must have a paid account (where you provide credit card details). For more information about creating a paid account, see Upgrading your account. To create a Plus plan Discovery service instance, go to the Discovery resource page in the IBM Cloud catalog and create a Plus plan service instance.

   > ⚠ **Important:** If you decide to stop using the Plus plan and don't want to pay for it, delete the Plus plan service instance before the 30-day trial period ends.

2. You should have access to the Discovery doc-tutorial-downloads repository to download the sample applications and data.

Additional prerequisites for the Granite application

1. Set up an instance of Watson Machine Learning. For more information about the pricing plans, see Watson Machine Learning.
2. Create an API key for IBM Cloud. For more information, see Understanding API keys.

Additional prerequisites for the Slate application

1. Set up an instance of Watson Machine Learning. For more information about the pricing plans, see Watson Machine Learning.
2. Set up an instance of Cloud Pak for Data 4.7.x or later, and install Watson Studio and Watson Machine Learning.
3. Create an API key for IBM Cloud. For more information, see Understanding API keys.
4. Create an API key for IBM Cloud Pak for Data. For more information, see Getting Started with IBM Cloud Paks.

Regex - Entity Extraction, document classification, and sentence classification by using regular expressions

In this sample, we are using IBM Cloud Code Engine as the infrastructure environment for the application of webhook enrichment. However, you can deploy the application in any other environment.

1. Deploy the webhook enrichment application to IBM Cloud Code Engine.
   1. Create a project in IBM Cloud Code Engine. For more information, see Create a project.
   2. Create a secret in the project. For more information, see Creating secrets.

      This secret must contain the following key-value pairs:

      - `WD_API_URL`: The API endpoint URL of your Discovery instance.
      - `WD_API_KEY`: The API key of your Discovery instance.
      - `WEBHOOK_SECRET`: A key to pass with the request that can be used to authenticate with the application. For example,`purple_unicorn`.

   3. Deploy the application from the sample repository source code. For more information, see Deploying your app from repository source code.

      In *Create application*, click **Specify build details** and enter these details.

      - For source, specify:
        - Code repo URL: URL of the sample code repository Discovery doc-tutorial-downloads page
        - Code repo access: `None`
        - Branch name: `master`
        - Context directory: `discovery-data/webhook-enrichment-sample/regex`
      - Strategy: `Dockerfile`
      - Output: Enter your container image registry information
      - Open **Environment variables (optional)**, and add the following environment variables:
        - Define as: `Reference to full secret`
        - Secret: The name of the secret that you created in the project in the previous step

You can set the **Min number of instances** to 1.

4. Ensure that the application status changes to **Ready**.

2. Configure the Discovery webhook enrichment. For more information, see Configuring the webhook enrichment.
3. Ingest documents to Discovery and see the results.
    1. Upload nhtsa.csv from Discovery doc-tutorial-downloads to the collection.
    2. Find the webhook enrichment results by previewing your query results after the document processing is complete.

Granite - Entity Extraction by using a foundation model of watsonx.ai

In this sample, we extract entities from an email by using watsonx.ai Granite model. We are using IBM Cloud Code Engine as the infrastructure environment for the application of webhook enrichment. However, you can deploy the application in any other environment.

1. Deploy the webhook enrichment application to IBM Cloud Code Engine.
    1. Create a project in IBM Cloud Code Engine. For more information, see Create a project.
    2. Create a secret in the project. For more information, see Creating secrets.

        This secret must contain the following key-value pairs:

        - `WD_API_URL`: The API endpoint URL of your Discovery instance.
        - `WD_API_KEY`: The API key of your Discovery instance.
        - `WEBHOOK_SECRET`: A key to pass with the request that can be used to authenticate with the application. For example,`purple_unicorn`.
        - `IBM_CLOUD_API_KEY`: The API key of IBM Cloud. It is used to access Watson Machine Learning API.
        - `WML_ENDPOINT_URL`: The API endpoint URL of your Watson Machine Learning. For more information, see the Machine Learning documentation.
        - `WML_INSTANCE_CRN`: The CRN of your Watson Machine Learning instance. You can find your instance and CRN using `ibmcloud` command: `ibmcloud resources`.
    3. Deploy the application from the sample repository source code. For more information, see Deploying your app from repository source code.

        In *Create application*, click **Specify build details** and enter these details.

        - For source, specify:
            - Code repo URL: URL of the sample code repository Discovery doc-tutorial-downloads page
            - Code repo access: `None`
            - Branch name: `master`
            - Context directory: `discovery-data/webhook-enrichment-sample/granite`
        - Strategy: `Dockerfile`
        - Output: Enter your container image registry information
        - Open **Environment variables (optional)**, and add the following environment variables:
            - Define as: `Reference to full secret`
            - Secret: The name of the secret that you created in the project in the previous step

        You can set the **Min number of instances** to 1.

    4. Ensure that the application status changes to **Ready**.
2. Configure the Discovery webhook enrichment. For more information, see Configuring the webhook enrichment.
3. Ingest documents to Discovery and see the results.
    1. Upload email.txt from Discovery doc-tutorial-downloads to the collection.
    2. Find the webhook enrichment results by previewing your query results after the document processing is complete.

Slate - Entity extraction with Watsonx.ai Slate model that is fine-tuned with labeled data exported from entity extractor workspace of Watson Discovery.

Slate models have the best cost performance trade-off for non-generative use cases. For fine tuning, it requires task-specific labeled data. You can prepare labeled data in Watson Discovery, fine-tune the Slate model in Watson Studio, and deploy the model in Watson Machine Learning. Once you deploy a fine-tuned model, you can create a webhook enrichment that enriches documents using that model in Watson Discovery.

1. Prepare labeled data in Watson Discovery.

1. Create an entity extractor workspace and label data. For more information, see Define custom entities.
2. Download labeled data from the entity extractor workspace. For more information, see Exporting labeled data for an entity extractor.

   In this tutorial, you can use the sample labled data from Discovery doc-tutorial-downloads in subsequent steps.

2. Fine tune the slate model in Watson Studio and deploy the model to Watson Machine Learning.
   1. Create a project in Watson Studio. For more information, see Creating a project.
   2. Create a deployment space in Watson Machine Learning. For more information, see Creating deployment spaces
   3. Create an environment template in the project. For more information, see Creating environment templates. You can create with the following options:
      - Type: `Default`
      - Hardware configuration
        - Reserve vCPU: 2
        - Reserve RAM (GB): 8
      - Software version: Runtime 23.1 on Python 3.10
   4. Create notebook in the project using the environment template as runtime from the notebook file. For more information about cxreating a notebook, see Creating notebooks. The notebook file is at Discovery doc-tutorial-downloads.
   5. Upload labeled data in the notebook. For more information, see Load data from local files.
   6. Fine tune and deploy the Slate model by running the notebook step-by-step and replacing certain variables.
3. Deploy the webhook enrichment application to IBM Cloud Code Engine.
   1. Create a project in IBM Cloud Code Engine. For more information, see Create a project.
   2. Create a secret in the project. For more information, see Creating secrets.

      This secret must contain the following key-value pairs:

      - `WD_API_URL`: The API endpoint URL of your Discovery instance.
      - `WD_API_KEY`: The API key of your Discovery instance.
      - `WEBHOOK_SECRET`: A key to pass with the request that can be used to authenticate with the application. For example,`purple_unicorn`.
      - `SCORING_API_HOSTNAME`: The API hostname of your Watson Machine Learning scoring deployment that serves your fine-tuned slate model.
      - `SCORING_DEPLOYMENT_ID`: The ID of your Watson Machine Learning scoring deployment that serves your fine-tuned slate model.
      - `SCORING_API_TOKEN`: The API token used in bearer authorization to use your Watson Machine Learning scoring deployment that serves your fine-tuned Slate model. You can get a token by using the following command:

```
curl -X POST {auth} \
SCORING_API_TOKEN=$(
curl -k -X POST 'https://{hostname of your cp4d instance}/icp4d-api/v1/authorize' \
                --header "Content-Type: application/json"
                -d "{\"username\":\"admin\",\"api_key\":\"{api key of your cp4d instance}\"}" \
| jq .token
)
```

4. Deploy the application from the sample repository source code. For more information, see Deploying your app from repository source code.
   1. In *Create application*, click **Specify build details** and enter these details.
      - For source, specify:
        - Code repo URL: URL of the sample code repository Discovery doc-tutorial-downloads page
        - Code repo access: `None`
        - Branch name: `master`
        - Context directory: `discovery-data/webhook-enrichment-sample/slate`
      - Strategy: `Dockerfile`
      - Output: Enter your container image registry information
      - Open **Environment variables (optional)**, and add the following environment variables:
        - Define as: `Reference to full secret`
        - Secret: The name of the secret that you created in the project in the previous step

      You can set the **Min number of instances** to 1.

2. Ensure that the application status changes to **Ready**.

5. Configure the Discovery webhook enrichment. For more information, see Configuring the webhook enrichment.

6. Ingest documents to Discovery and see the results.

    1. Upload a page of Annual report from Discovery doc-tutorial-downloads to the collection.
    2. Find the webhook enrichment results by previewing your query results after the document processing is complete.

Configuring the webhook enrichment

1. Create a project.

2. Create a webhook enrichment by using the Discovery API.

```
curl -X POST {auth} \
--header 'Content-Type: multipart/form-data' \
--form 'enrichment={"name":"my-first-webhook-enrichment", \
"type":"webhook", \
"options":{"url":"{your_code_engine_app_domain}/webhook", \
   "secret":"{your_webhook_secret}", \
   "location_encoding":"utf-32"}}' \
'{url}/v2/projects/{project_id}/enrichments?version=2023-03-31'
```

3. Create a collection in the project and apply the webhook enrichment to the collection.

```
curl -X POST {auth} \
--header 'Content-Type: application/json' \
--data '{"name":"my-collection", \
"enrichments":[{"enrichment_id":"{enrichment_id}", \
   "fields":["text"]}]}' \
'{url}/v2/projects/{project_id}/collections?version=2023-03-31'
```

Start getting value from your data

Learn what IBM Watson® Discovery can do to help you find answers, recognize patterns, and gain insights from your data.

Find checklists of the high-level steps to follow to achieve the following goals:

- Pinpoint answers
- Extract meaning
- Enhance your chatbot
- Find trends
- Analyze contracts

Pinpoint answers

Help customers find answers faster. Analyze content from various connected data sources, pinpoint the most relevant passage or phrase, and return the right information when someone asks for it.

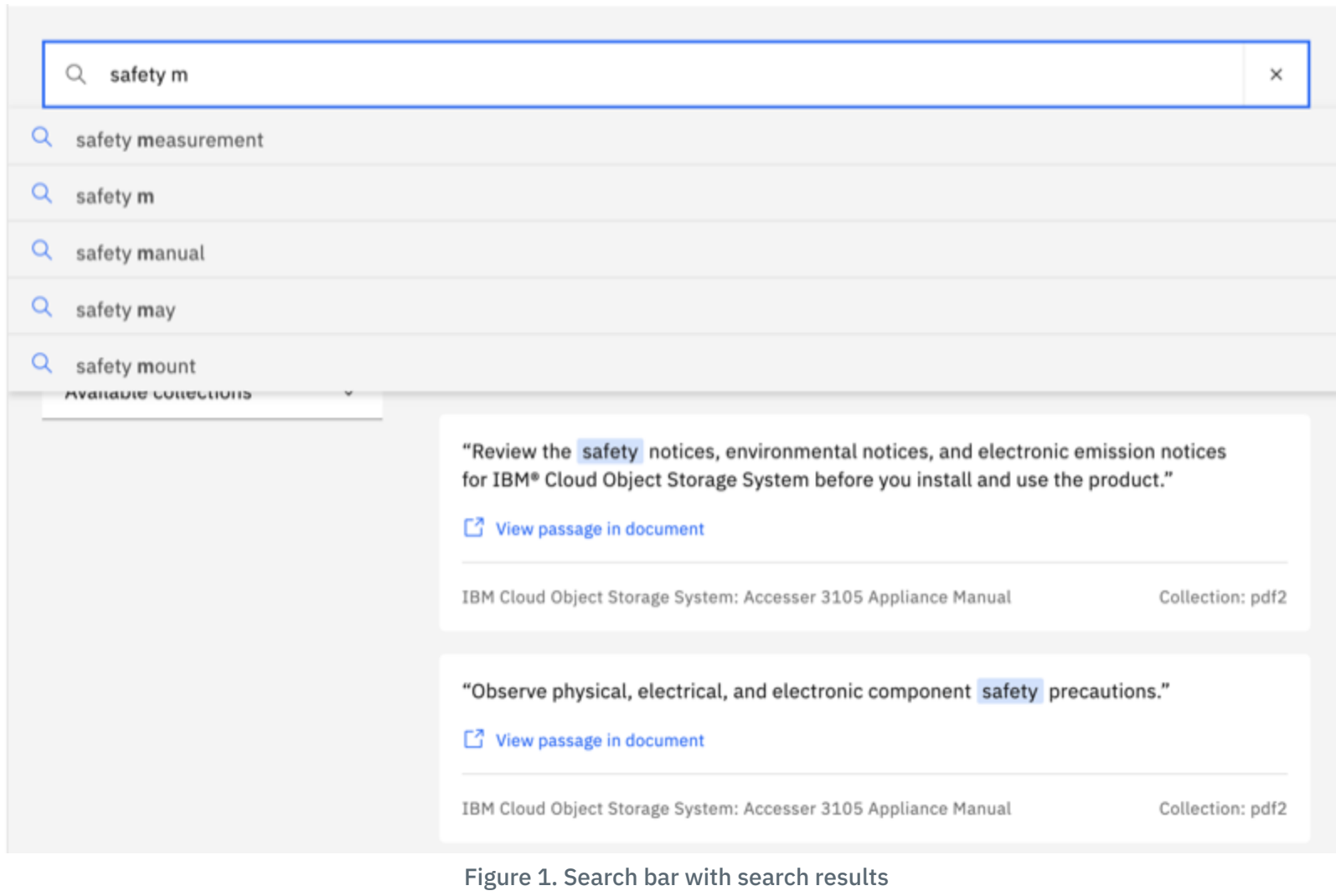

Figure 1. Search bar with search results

If pinpointing answers is your goal, complete the steps that are listed in the following table.

Step	Task	Related information
☐	Create a *Document Retrieval* project.	Creating projects
☐	Add up to 5 collections that connect to external data sources or contain uploaded files.	Creating collections
☐	Run test queries to assess the quality of the initial results.	Previewing the default query results
☐	Take actions to improve your results. For example, you can customize the search bar to enable autocompletion.	Improving your query results
☐	Deploy your search solution.	Deploying your project

Checklist for getting answers

Extract meaning

Use award-winning natural language processing technology to enrich your data and ensure that the right information is found when someone searches for answers.

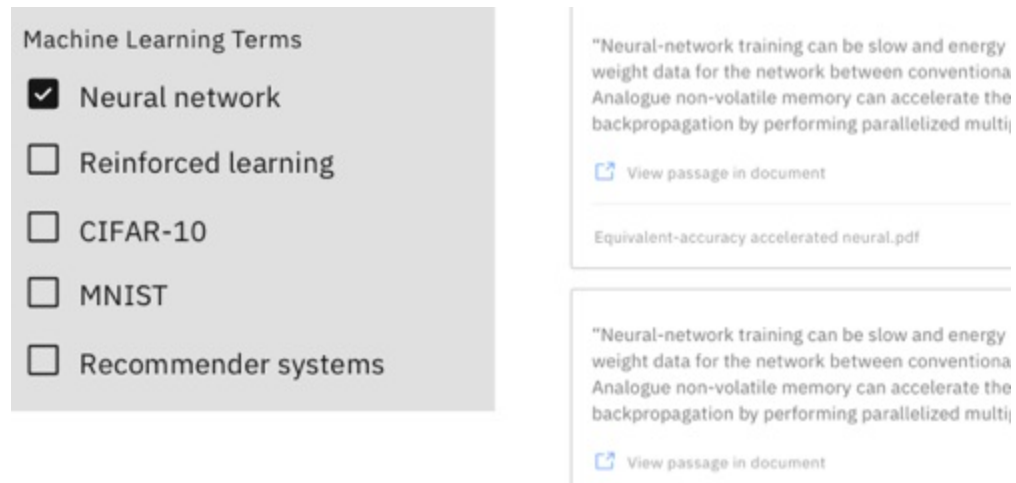

Figure 2. Machine learning facet for filtering search results with custom enrichment values

If extracting meaning is your goal, complete the steps that are listed in the following table.

Step	Task	Related information
☐	Create any project type.	Creating projects
☐	Add collections that connect to external data sources or contain uploaded files.	Creating collections
☐	Chunk large documents into many smaller documents so you can apply more targeted enrichments to the content.	Using Smart Document Understanding
☐	Enhance your data by applying built-in NLU enrichments.	Applying prebuilt enrichments
☐	Identify and promote terms and patterns from your data with special significance to your use case.	Adding domain-specific resources
☐	Submit test queries to assess the results.	Testing your project
☐	Create a facet that surfaces the enriched data from your documents.	Facets
☐	Take actions to improve your results.	Improving your query results
☐	Deploy your solution.	Deploying your project

Checklist for extracting meaning

Enhance your chatbot

Delight your customers by fortifying your chatbot with an answer to every question. Discovery is designed to work seemlessly with watsonx Assistant to search and deliver answers from help content that you already own.

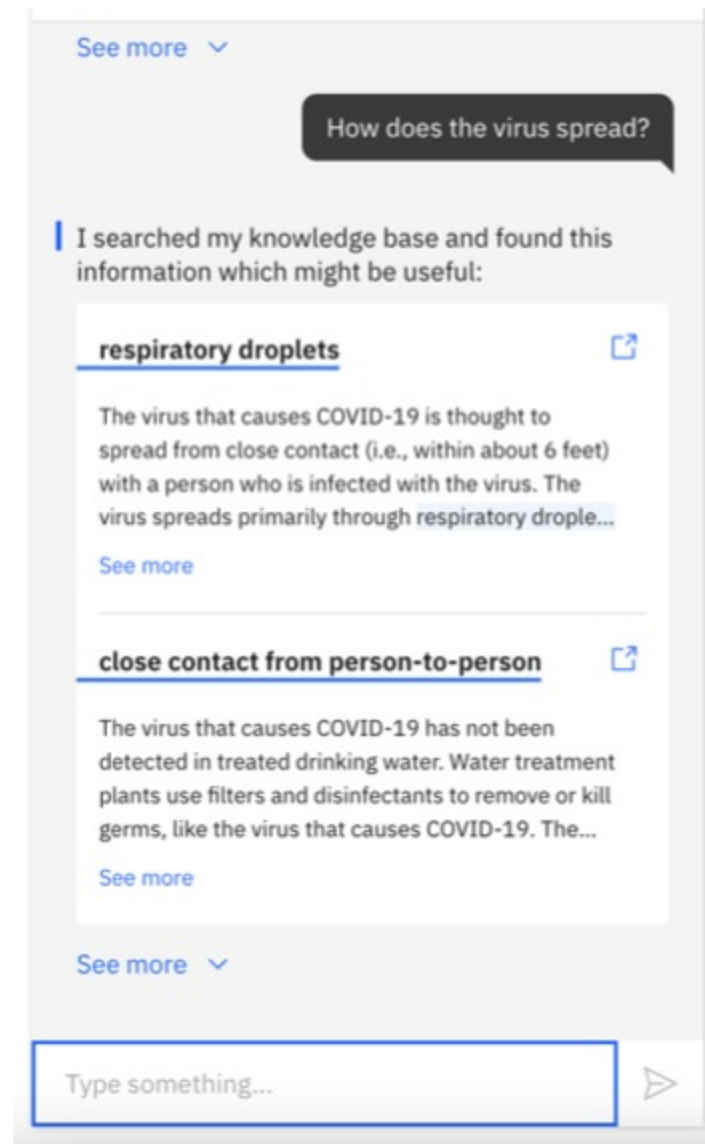

Figure 3. Answer finding enabled in the watsonx Assistant web chat

If enhancing your chatbot is your goal, complete the steps that are listed in the following table.

Step	Task	Related information
☐	Create a *Conversational Search* project.	Creating projects
☐	Add a collection that connects to an external data source or contains uploaded files.	Creating collections
☐	Run test queries to assess the quality of the initial results.	Previewing the default query results
☐	Take actions to improve your results.	Improving your query results
☐	Connect your project to a virtual assistant that is built with Watson Assistant.	Deploying your project
☐	From the Watson Assistant user interface, deploy the web chat that is associated with your assistant.	Deploying your assistant

Checklist for enhancing your chatbot

For a more detailed look at these steps, take a tutorial that walks you through them. For more information, see Power your assistant with answers from web resources.

Alternatively, you can add a generative language service named NeuralSeek between the Watson Discovery and watsonx Assistant services. For more information, see Use NeuralSeek to return polished answers from existing help content .

Find trends

Uncover patterns, trends, and relationships in structured and unstructured data. Use text analytics to gain insights into social media, e-commerce trends, and user behavior. Or start to address problems by finding their root cause.

Note: Only users of installed deployments (IBM Cloud Pak for Data) or Premium or Enterprise plan-managed deployments can create this type of project.

Figure 4. Analyzing data with the Content Mining application

If finding trends in your data is your goal, complete the steps that are listed in the following table.

Step	Task	Related information
☐	Create a *Content Mining* project.	Creating projects
☐	Add a collection that connects to an external data source or contains uploaded files.	Creating collections
☐	Use the built-in Content mining application to analyze your data.	Analyzing your data with the content mining application

Checklist for finding trends

Analyze contracts

Accelerate the pace at which experts can analyze complex documents.

Note: Only users of installed deployments (IBM Cloud Pak for Data) or Premium or Enterprise plan-managed deployments can create this type of project.

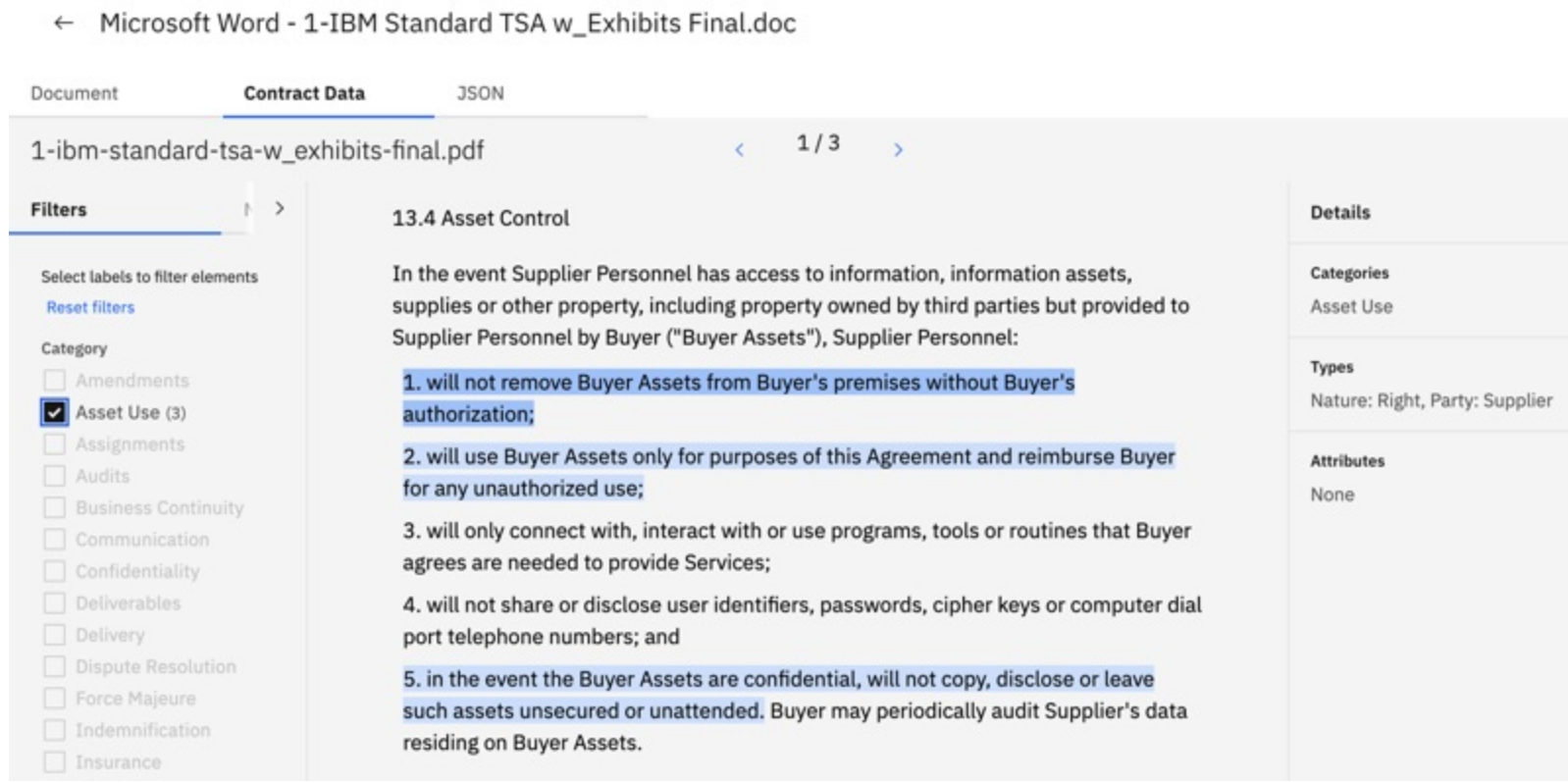

Figure 5. Analyzing contracts

If analyzing contracts is your goal, complete the steps that are listed in the following table.

Step	Task	Related information
☐	Create a *Document Retrieval for Contracts* project.	Creating projects
☐	Add up to 5 collections that connect to external data sources or contain uploaded files.	Creating collections
☐	Run test queries to assess the quality of the initial results.	Previewing the default query results
☐	Take actions to improve your results.	Improving your query results
☐	Analyze the data.	Understanding contracts

Checklist for analyzing contracts

Connecting to your data

Creating projects

A project is a convenient way to collect and manage the resources in your IBM Watson® Discovery application. You can assign a *Project type* and connect your data to the project by creating a collection.

Before you create a project, decide which project type best fits your needs.

Project descriptions

Need	Goal	Project type
Which document contains the answer to my question?	Find meaningful information in sources that contain a mix of structured and unstructured data, and surface it in a stand-alone enterprise search application or in the search field of a business application.	**Document Retrieval**
Where is the part of the contract that I need for my task?	Quickly extract critical information from contracts.	**Document Retrieval for Contracts**
I want the chatbot I'm building to use knowledge that I own.	Give a virtual assistant quick access to technical information that is stored in various external data sources and document formats to answer customer questions.	**Conversational Search**
I want to uncover insights I didn't know to ask about.	Gain insights from pattern analysis or perform root cause analysis.	**Content Mining**

Project type use cases

Note: If you created the Discovery service as part of a IBM Cloud Pak for Data as a Service deployment, the Discovery project is separate and distinct from the deployment project that is displayed in IBM Cloud.

To create a project, complete the following steps:

1. Open the *Projects* page by selecting **My Projects**.
2. Click **New project**. Name your project, and then choose the project type.

   For more information about each type, see Project types.

   Otherwise, choose **None of the above** and a *Custom* project type is created for you.
3. If you choose a *Document Retrieval* project type and your data sources are in English, decide whether to enable the Content Intelligence feature.

   If your data source contains contracts, enable the feature by selecting **Apply contracts enrichment**. Scroll to see the checkbox, if necessary.
4. Click **Next**.
5. Choose and configure a data source or connect to an existing collection.

   For more information about supported data sources, see Creating collections.

Take advantage of the following resources that are available from the page header:

- To open the product documentation, click the Help icon .
- To see all of your projects, click **My projects**.

Project types

Choose a project type to get the correct set of enrichments applied to your documents automatically. The improvement tools that are available differ by project type, as do the deployment methods, which are optimized for each use case.

The following project types are available:

- Document Retrieval
- Document Retrieval for Contracts
- Conversational Search
- Content Mining

- Custom

For more information about the different settings that are applied to each project type, see Default project settings.

Document Retrieval

Use this project type to search and find the most relevant answers from your data. Projects of this type are typically deployed as search field components that are added to websites or other applications.

Documents that you add to a project of this type are automatically enriched in the following ways:

- Entities, such as proper nouns, are identified and tagged.
- Parts of speech are identified and tagged.

This tagged information is used later when a natural language phrase is submitted as a search query to return a smarter response.

Tip: A sample Document Retrieval project is available for you to explore. For more information, see Getting started with Watson Discovery.

Document Retrieval for Contracts

If you are working with English-language legal contracts, enable the Content Intelligence feature to apply a contracts enrichment that can recognize and tag contract-related concepts in your data. Use this project type to automate complex business processes, such as contract review and negotiation. This project type can help to increase productivity, minimize costs, and reduce your legal exposure.

Note: Only users of installed deployments (IBM Cloud Pak for Data) or Premium or Enterprise plan managed deployments can create this type of project.

In addition to the enrichments that are applied to a typical document retrieval project, the following enrichments are made automatically:

- Content from tables in the source document is tagged so that it can be found later.
- Contract details, such as payment terms or parties that are involved in the contract, are identified and tagged.

For any collection that you add to the project, optical character recognition (OCR) is enabled automatically so that text from scanned documents or other images is processed.

Note: When you apply the contracts enrichment, you cannot use Smart Document Understanding to annotate documents. A pretrained SDU model that can recognize contract-related information is applied automatically. The Table understanding enrichment is automatically applied.

For more information, see Understanding contracts.

Conversational Search

The *Conversational Search* project returns information from a connected data collection as answers to questions that customers ask a chatbot, which is also known as an *assistant*.

Use IBM® watsonx™ Assistant and Discovery together to give your assistant access to technical content and other knowledge base resources without having to relocate or copy your corporate data. The built-in synchronization capabilities mean that your assistant can share the most up-to-date information available. Use the integrations that are provided with watsonx Assistant to deploy an assistant that connects to this project to various platforms, including your company website, in minutes.

The documents that you add to this type of project are not enriched automatically.

If you need to perform more complex searches from your virtual assistant, you might want to create a *Document Retrieval* project instead of *Conversational Search* project. For more information, see Choosing the right project type for a chatbot.

IBM Cloud Another feature to consider enabling is the *Emphasize the answer* feature. When enabled, the answers that are returned to customers who interact with the assistant show the exact answer highlighted in bold font within the search response. For more information about how the exact answer is determined, see Answer finding.

For more information about building a watsonx Assistant search skill, see the appropriate documentation for your deployment:

- IBM Cloud Pak for Data Adding a search integration.
- IBM Cloud Embedding existing help content

Note: From the classic watsonx Assistant experience, see Creating a search skill.

Content Mining

Use this project type to discover hidden insights, trends, and relationships in your data.

Note: Only users of installed deployments (IBM Cloud Pak for Data) or Premium or Enterprise plan managed deployments can create this type of project.

This project type is especially useful for analyzing structured data, such as data that you add by uploading a CSV file or by connecting to a database data source. You can add only one collection to a project of this type from the Discovery user interface.

Documents that you add as part of the initial collection are automatically enriched in the following way:

- Parts of speech are identified and tagged.

After you add a collection and optionally apply more enrichments to the data, a full-featured application is available for you to deploy. You can use the application to research your data in depth. For more information about using the application, see Analyzing your data with the deployed Content Mining application.

From the Content Mining application, you can create the following enrichment types which are not available in other project types:

- Document classifier
- Phrase sentiment

Note: You can create a collection from the deployed Content Mining application. The collection that you create is not added to your existing Content Mining project. A new Content Mining project is created to store the collection. The collection can contain an uploaded CSV file only. The project that is generated is given the name that you specify for the collection.

Because the data that you add to this type of project is often structured, consider using the API to submit queries in the Discovery Query Language (DQL). With DQL queries, you can get information from specific fields or find specific enrichment type mentions. You cannot apply relevancy training to a *Content Mining* project.

Custom

Choose this type if you prefer not to use one of the other project types. No enrichments are applied automatically, so you can add only those enrichments that are necessary for your use case.

Basic project defaults

Some enrichments and query result settings are applied to each project type by default.

Project type	Default enrichments	Default query result settings
Document Retrieval	Entities	Facets (by Entity), Passages
Document Retrieval for Contracts	Entities, Table Understanding, and Contracts	Facets (by Category, Nature, Contract Term, Contract Payment Term, Contract Type, Contract Currency, Invoice Buyer, Invoice supplier, Invoice Currency, Purchase Order Buyer, Purchase Order Supplier, Purchase Order Payment Term) and Table Retrieval
Conversational Search	None	Passages
Content Mining	Part of Speech	None
Custom	None	Passages

Basic project defaults

Project limits

The number of projects you can create depends on your Discovery plan type.

Plan	Projects per service instance
Cloud Pak for Data	Unlimited

Premium	100
Enterprise	100
Plus (includes Trial)	20

Plan details

The Sample project is excluded from the total number of projects.

Renaming a project

Note: You cannot rename the *Sample Project*.

To rename a project after you create it, complete the following steps:

1. Go to the *My Projects* page.
2. Find the project that you want to rename, click the *Project actions* icon ⋮, and then choose **Rename**.
3. Edit the project name, and then click **Apply**.

Deleting a project

If you want to delete a project, but keep a collection from the project, share the collection with another project before you complete these steps. From another project (a type that allows multiple collections), open the *Manage collections* tab. Click **New collection**, and then click **Reuse data from an existing collection**. Select the collection that you want to keep, and then click **Finish**.

Note: You cannot delete the *Sample Project*.

To delete a project, complete the following steps:

1. Go to the *My Projects* page.
2. Find the project that you want to delete, click the *Project actions* icon ⋮, and then choose **Delete**.
3. Click **Delete**.

Creating collections

A collection is a set of documents that you add to a project so that you can analyze, enrich, and extract useful information from it.

You can add data to your project in the following ways:

- Upload locally accessible files by using the product user interface. This method is the best way to get started and test your use case.
- Set up a scheduled crawl of documents that are stored on an external data source.

  The product user interface offers several built-in data source connectors for you to choose from. The options differ depending on your deployment type. For more information, see Supported data sources.
- Connect to an external data source for which no built-in support is available:

  IBM Cloud

  Use IBM App Connect to set up a scheduled crawl of documents that are stored on other external data sources.

  IBM Cloud Pak for Data

  Build a connector to crawl documents that are stored on other external data sources.
- To automate the process of adding data to your project, use the Discovery APIs to create a collection and upload documents to it.

When you add documents to Discovery, the original documents are crawled and information from the documents is stored in an index so that it can be enriched and analyzed or retrieved later. Not all rich content from the original document is retained. For example, images from .ppt or .doc files are not stored. For more information, see How your data source is processed.

IBM Cloud After you create a collection, you can click **Preview data** to preview data in the advanced document view.

Choosing what to add to a collection

There a few things to consider as you decide how to break up your source content into collections.

- Getting content from different data sources

  If you store similar content in more than one type of data source (a website and Salesforce, for example), you can create one project with two separate collections. Each collection adds documents from a single data source. When they are built together into a single project, a user can search across both sources at the same time.

- Applying enrichments

  Creating a collection is a good way to group documents that you want to enrich in a similar way. For example, maybe a subset of your documents contains industry jargon and you want to add a dictionary that recognizes the terms. You can create a separate collection and use the term suggestions feature to speed up the process of creating the dictionary.

- Creating separate Smart Document Understanding (SDU) models

  You can use the Smart Document Understanding tool to identify content based on the structure of a document. If you have 20 PDF files that were created by your Sales department and use one template and 20 PDF files that were created by your Research department and use a different template, group each set into its own collection. You can then use the SDU tool to build a model for each structure separately, a model that understands the unique structure. You can also use the tool to define custom fields that are unique to the source documents.

Creating a collection

Before you can create a collection, you must create a project. For more information, see Creating projects.

Things to keep in mind:

- A collection can support only one external data source.
- Documents in the collection must be in one language only, the language that you specify for the collection.

To create a collection, complete the following steps:

1. Open a project, go to the *Manage collections* page, and then click **New collection**.
   - The Conversational Search, Document Retrieval, and Custom project types can contain up to 5 collections.
   - A Content Mining project can contain only 1 collection.
2. Choose how you want to add data to your collection.
   - Uploading data
   - Reusing data from a collection
   - Crawling an external data source.

     For supported data sources, see the appropriate topic for your deployment type:

     - IBM Cloud Pak for Data IBM Cloud Pak for Data data sources
     - IBM Cloud IBM Cloud data sources

     Tip: These topics also describe how to connect to data sources that are not supported by default per deployment type.

For information about how to troubleshoot issues that you might encounter when you add documents to a collection, see Troubleshooting ingestion.

For more information about how to create a collection programmatically, see the API reference documentation.

Optical character recognition

One of the optional features that you can apply to a collection when you create it is optical character recognition. The optical character recognition (OCR) feature extracts text from images. This capability is useful for preserving information that is depicted in diagrams or graphs, or in text that is embedded in files such as scanned PDFs. By converting the visual information into text, it can later be searched.

A new version of the technology was introduced in cloud-managed instances. OCR v2 was developed by IBM Research to be better at extracting text from scanned documents and other images that have the following limitations:

- Low-quality images due to incorrect scanner settings, insufficient resolution, bad lighting (such as with mobile capture), loss of focus, misaligned pages, and badly printed documents
- Documents with irregular fonts or various colors, font sizes, and backgrounds

Things to keep in mind when you enable OCR:

- The time that it takes to ingest a document with images increases when OCR is enabled.
- OCR can read both clear and noisy images. It can convert noisy images to gray scale, and smooth and de-skew them. However, the image quality

must meet the minimum requirement of **80 DPI** (dots per inch).
- OCR can recognize many languages, but the language of the text in the image must be the same as the language that is specified for the collection where the file is added.

For more information about languages for which OCR v1 and OCR v2 are supported, see Language support.

For a list of files types where you can apply OCR, see the Supported file types table.

Enabling stemming for uncurated data IBM Cloud Pak for Data

Note: This feature is available from IBM Cloud Pak for Data deployments only. It was introduced with the 4.7.0 release.

You can configure Discovery to use stemming instead of lemmatization for normalization when you create a collection. This configuration is only occasionally useful when collections, queries, or both contain data with many misspellings, missing accent marks, and grammatical errors.

Discovery normalizes words to enable faster recognition and matching of words and their various forms, such as plurals or alternative verb conjugations. By default, Discovery uses lemmatization to normalize words based on their meaning. Stemming normalizes words by using word stems only.

Lemmatization is more precise, but works best on curated data. If your data is not well curated, stemming might work better. The same word stem typically is detected whether or not a word is spelled correctly. However, lemmatization might not recognize a misspelled word or might misinterpret its meaning. As a result, the lemmatizer can add the wrong root word to represent the misspelled word in the index. A search against a stemmed version of a misspelled word is likely to return better results than a search against an incorrectly lemmatized word.

The following table shows examples of how some words are stemmed versus lemmatized.

Surface form	Lemmatized form	Stemmed form
running	run	run
ran	run	ran
instructor	instructor	instruct
instruction	instruction	instruct

Stemmer versus Lemmatizer comparison

As you can see from the examples, the lemmatizer captures the word meanings better than the stemmer. Both *running* and *ran* are recognized as different forms of the same root verb *run*. And the difference in meaning between the two nouns *instructor* and *instruction* is preserved. However, if the data contains misspellings such as *instructer* and *instructoin*, the normalized form that is generated by stemming (*instruct*) will return better matches.

Discovery normalizes words when it ingests and stores data in the index and at run time when it analyzes queries that are submitted by users. The same normalization method is used for both operations, even though one operation occurs at the collection-level and the other occurs at the project-level. When a query is submitted, it is federated to each collection within the project, where the query is normalized based on that collection's configuration. Collections that are configured to use the stemmer normalize the query by using stemming. The collections that are not, normalize the query by using lemmatization.

To enable the stemmer instead of the lemmatizer when you create the collection, expand **More processing options**, and then set the *Use stemming instead of lemmatization when indexing* switcher to **On**.

If you configure Discovery to use the stemmer, consider also designing the queries that extract information from the collection to allow for character differences during matching. For more information, see the String variation operator.

For more information about the languages for which the stemmer is supported, see Language support.

Collection limits

The number of collections that you can create per project differs by project type.

Project type	Collections per project
Document Retrieval	5
Document Retrieval for Contracts	5
Conversational Search	5

Content Mining	1
Custom	5

Collections per project limits

The number of collections you can create per service instance depends on your Discovery plan type.

Plan	Collections per service instance
Cloud Pak for Data	300
Premium	300
Enterprise	300
Plus (includes Trial)	40

Plan details

IBM Cloud Pak for Data The number of collections you can create depends on your hardware configuration. Discovery supports a maximum of 300 collections per instance and installation, but that number depends on many factors, including memory.

Supported file types

Discovery can ingest specific file types. For all other types of files, a warning message is displayed and the file is not ingested.

The following table shows the supported file types and information about feature support that varies by file type.

File type	Text extraction support	Smart Document Understanding (SDU) support	Optical Character Recognition (OCR) support
CSV	✓		
DOC, DOCX	✓	✓	✓
GIF	✓		
HTML	✓		
JPG	✓	✓	✓
JSON	✓		
PDF	✓	✓	✓
PNG	✓	✓	✓
PPT, PPTX	✓	✓	✓
TIFF	✓	✓	✓
TXT	✓		
XLS, XLSX	✓		✓

Supported file types

- PDF files that are secured with a password or certificate are not supported. Vector objects, including SVG images and vectorized text, are not supported. Only images of the supported image file types that occur in the PDF are rendered.
- Only single-page image files are supported.
- Files within compressed archive files (ZIP, GZIP, TAR) are extracted. Discovery ingests the supported file types within the archive; it ignores all other file types. The file names must be encoded in UTF-8. Files with names that include Japanese characters, for example, must be renamed before they

are added to the ZIP file.
- Discovery supports MacOS ZIP files only if they are generated by using a command such as: `zip -r my-folder.zip my-folder -x "*.DS_Store"`. ZIP files that are created by right-clicking a folder and clicking *Compress* are not supported.
- PDF files that you upload as part of an archive file are not displayed in the advanced view for a query result that you open from the *Improve and customize* page. If you want the file to be viewable from the advanced view, reimport the PDF file separately from the archive file.

Note: When you add files to a Document Retrieval for Contracts project type, any file types that support SDU and OCR are processed with a pretrained Smart Document Understanding model and Optical Character Recognition automatically.

Document limits

The number of documents that are allowed per service instance depends on your Discovery plan type.

The document limit applies to the number of documents in the index. Upload fewer documents at the start if the enrichments that you plan to apply might increase the number of documents later. For example, the following configurations generate more documents:

- When you split a document, the document is segmented into multiple documents
- CSV files that you upload generate one document per line
- Database data sources that you crawl produce one document per database row
- Each object that is defined in an array in a JSON file results in a separate document

Plan	Documents per service instance
Cloud Pak for Data	Unlimited
Premium	Unlimited
Enterprise	Unlimited
Plus (includes Trial)	500,000

Number of documents per service instance

For the Enterprise plan, you are charged after 100,000 documents per month. For more information about pricing, see Discovery pricing plans.

Note: The maximum allowed number can vary slightly depending on the size of the documents. Use these values as a general guideline.

File size limits

Crawled documents

The maximum size of each file that you can crawl by using a connector differs by deployment type.

IBM Cloud Managed deployments on IBM Cloud

- Premium plans only:
  - Box: 50 MB
  - IBM Cloud Object Store: 50 MB
  - Salesforce Files objects: 50 MB
  - All other data sources: 10 MB
- All other plans: 10 MB

IBM Cloud Pak for Data Installed deployments on IBM Cloud Pak for Data

- All data sources: 32 MB

Uploaded documents

The size of each file that you can upload depends on your Discovery plan type. See the *Maximum document size table for details.

Plan	File size per document
Cloud Pak for Data	50 MB

Premium	50 MB
Enterprise	10 MB
Plus (includes Trial)	10 MB

Maximum document size

Field limits

When a document is added to a collection, content from the document is evaluated and added to the appropriate fields in an internal index.

For structured data, such as uploaded CSV or JSON files, or data from crawled databases, each column or object is stored as a root-level field. For example, if you add a CSV file to collection, each column in the CSV file is stored as a separate field in the index.

A maximum of 1,000 fields can be added to the index.

You cannot assign the data type, such as Date or String, of a field. The data type is detected automatically and assigned to the field during document ingestion. The assignment is based on the data type that is detected from the first document that is indexed. Ingestion errors can occur in subsequent documents if a different data type is detected for the value in the same field. Therefore, if your documents have a mix of data types in a single field, first ingest the document that has a value with the most flexible data type, such as String, in the field.

When you crawl a website or upload an HTML file, the HTML content is added to the collection and indexed in an `html` field.

The following table shows the maximum size limit for fields per document.

Field type	Maximum allowed size per document
`html` field	5 MB
Sum of all other fields	1 MB

Maximum field sizes

If the maximum size of the fields in the document exceeds the allowed limits, they are treated as follows:

- For a document with an oversized `html` field, all of the fields in the document are indexed except the `html` field.

  Note: For IBM Cloud Pak for Data version 4.0 and earlier, the entire document is not indexed.

- For a document with oversized non-HTML fields, the document is not indexed.

Tip: If you are uploading a Microsoft Excel file and a message is displayed that indicates that the non-HTML field size limit is exceeded, consider converting the XLS file into a CSV file. When you upload a comma-separated value (CSV) file, each row is indexed as a separate document. As a result, no field size limits are exceeded.

For more information about how fields in uploaded files are handled, see How fields are handled.

Supported data sources

The following table shows the supported data sources for each deployment type.

Data source	IBM Cloud	IBM Cloud Pak for Data
Box	✓	✓
Database (IBM Data Virtualization, IBM Db2, Microsoft SQL, Oracle, Postgres)		✓
FileNet P8		✓
HCL Notes		✓
IBM Cloud Object Storage	✓	

Local file system		✓
Salesforce	✓	✓
Microsoft SharePoint Online	✓	✓
Microsoft SharePoint On Premises	✓	✓
Website	✓	✓
Microsoft Windows file system		✓

Supported data sources

Crawl schedule options

When you create a collection, the initial crawl starts immediately. The frequency that you choose for the crawl schedule determines when the next crawl will start.

To create a crawl schedule, complete the following steps:

1. In the *Crawl schedule* section, choose a frequency.

   You can schedule the crawler to run at a specific day and time. This option is helpful if you want to avoid heavy load on a target system during business hours. If you specify an hour in the range 1 - 9, add a zero before the hour digit. For example, you can schedule the crawl for `01:00 AM` on Saturdays.

   IBM Cloud When you schedule a crawl to run monthly, the day number options are limited to 1 through 28 because you must specify a day that occurs every month, including February which has 28 days.

   IBM Cloud Pak for Data Installed deployments have more schedule options:

   - If you want to crawl every 12 hours or every 10 days, choose **Custom intervals**. You can schedule the crawler to run on a custom number of days or hours.
   - By default, the crawl is scheduled to start during off-peak hours.
   - Do not set the interval to a frequency that is shorter than the time it takes for the crawl to finish.
   - Do not configure multiple crawlers to run at short intervals.
   - If you open a collection in a time zone other than the one in which the collection was created, the Coordinated Universal Time (UTC) offset information is displayed.

2. IBM Cloud Pak for Data Installed deployments have a **More scheduling settings** section where you can choose the type of schedule to use to crawl the data source.

   The choices for all of the connectors (except the *Web crawl* connector) are as follows:

   - **Full crawling**: Recrawls the external data source to update documents in the collection.
   - **Crawling updates (look for new, modified, and deleted contents)**: Updates the collection only if data in the external data source was added, modified, or deleted since the last crawl.
   - **Crawling new and modified contents**: Updates the collection only if data in the external data source that was added or modified since the last crawl.

Web crawl connector only: The *Web crawl* connector schedules crawls differently from the other connector types. For the *Web crawl* connector only, choose among the following options:

- To control the frequency of the crawls yourself, choose this option:

  Full crawling

  When you choose a full crawl schedule type, the crawl occurs with the frequency that you specify in the *Crawl schedule* section of the page.

- To allow the system to manage the frequency of the crawls for you, choose one of the following options:

  Crawling updates (look for new, modified, and deleted contents) or **Crawling new and modified contents**

  When you choose a schedule type that crawls for updates or for new and modified contents, the frequency that you specify for the crawl schedule is ignored. The frequency with which each document is crawled is variable and is managed entirely by the service. And the frequency changes depending on how often changes are found in a document. For example, if 5 of the 10 documents in a collection changed by the end of the first crawl interval, then the frequency is automatically increased for those 5 documents. Currently, the highest frequency at which these self-managed refreshes can run is daily.

  You cannot interrupt the automated management of frequency and you cannot trigger a one-off crawl when these types of scheduled crawls

are configured.

If you want to change the flexible crawl schedule settings later, you can go to the *Processing settings* page, edit the settings, and then click **Apply changes and reprocess**.

IBM Cloud The next scheduled crawl is displayed on the Activity page.

If you change the schedule frequency, the next scheduled crawl time might not be what you expect. The crawls are set up to occur on a regular schedule at a specific time or day by default. For example, if you change the crawl schedule from weekly to monthly on 11 August, the next crawl might be scheduled for 31 August instead of 11 September. It is not scheduled for exactly a month from the day that you made the change. Instead, it is scheduled to run on the day that is designated as the default run day for the selected crawl frequency.

Stopping a crawl

You can stop a crawl without changing the crawl schedule frequency. This action is helpful if you want to perform a time-consuming task and do not want the crawl to start or run in between the task.

IBM Cloud To stop a crawl, complete the following steps:

1. Open the *Manage collections* page from the navigation panel.
2. Select the collection for which you want to stop the crawl.
3. On the *Activity* page, if the crawl is in progress, click **Stop**.
4. Go to the *Processing settings* page.
5. Set **Apply Schedule** to **No**, and then click **Apply changes and reprocess**.

   The crawl is stopped and will not start again until you restart it.

IBM Cloud To restart the crawl, complete the following steps:

1. Open the *Manage collections* page from the navigation panel.
2. Select the collection for which you want to restart the crawl.
3. Go to the *Processing settings* page.
4. Set **Apply Schedule** to **Yes**, and then click **Apply changes and reprocess**.

   The crawl starts immediately.

   The next crawl will start based on the frequency that is selected in the crawl schedule options. If you want to start the crawl at any time before the scheduled frequency, click **Recrawl** on the *Activity* page.

IBM Cloud Pak for Data

You can temporarily stop a crawl that is in progress.

To stop a crawl temporarily, complete the following steps:

1. Open the *Manage collections* page from the navigation panel.
2. Select the collection for which you want to stop the crawl temporarily.
3. On the *Activity* page, click **Stop**.

   The crawl starts again based on the frequency that is specified in the crawl schedule.

Uploading data

You can perform a one-time document upload from your local file system at any time to add data to a project.

You can upload up to 200 files at a time.

To process document sets that are larger than 200 files, you can add them to an external data source and use a data source crawler to upload them. For IBM Cloud Pak for Data deployments, you can use a *Local File System* data source for this purpose.

For more information about the maximum size allowed for each file, see Document limits.

Tip: Before you upload a CSV file to a Content Mining project, consider adding headers to the source file so that any fields that are generated from the file have meaningful names. Without headers, fields are given generic names, such as `column_0`, `column_1`, and so on.

To upload data, complete the following steps:

1. Open your project, go to the **Manage collections** page, and then click **New collection**.
2. Choose **Upload data** as your data source, and then click **Next**.
3. Name the collection.
4. If the language of the documents in storage is not English, select the appropriate language.

   For a list of supported languages, see Language support.
5. Optionally, click **More processing settings** to expand the menu, and then click **Apply optical character recognition (OCR)**.

   > **Note:** When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.
6. Click **Next**.
7. Browse for the files you want to crawl.

   IBM Cloud You can drag documents that you want to add to your collection.

   For more information about supported file types, see Supported file types.
8. Click **Finish**.

The file upload is completed quickly. It takes more time for the data to be processed as it is added to the collection. After the files are uploaded and processed, the *Activity* page shows the upload results.

Unlike crawled data sources, you cannot schedule regular updates for uploaded files. If you want to add a later version of a file, delete the earlier version of the file, and then upload the latest version.

For information about how to troubleshoot issues that you might encounter when adding documents to a collection, see Troubleshooting ingestion.

For more information about what happens next, see How your data source is processed.

Configuring IBM Cloud data sources

Overview of IBM Cloud data sources

You can use IBM Watson® Discovery on the IBM Cloud® to connect to and crawl documents from remote sources.

IBM Cloud **IBM Cloud only**

> **Note:** This information applies only to managed deployments. For more information about IBM Cloud Pak for Data data sources, see Overview of Cloud Pak for Data data sources.

Connect to an external data source so that you can pull documents into Discovery on a schedule. Discovery pulls documents from the data source by *crawling* the data source. Crawling is the process of systematically browsing and retrieving documents from a starting location that you specify. When the crawler first processes a data source, it performs a full crawl. Each time the crawler runs after the initial crawl, it performs a refresh, where it checks for new and changed files only.

> **Important:** All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

You can use Discovery to crawl from the following data sources:

- Box
- IBM Cloud Object Storage
- Microsoft SharePoint Online
- Microsoft SharePoint On Prem
- Salesforce
- Web crawl

Your data source isn't listed? Check whether IBM® App Connect has a connector to the data source. You can use a default connector that is built for App Connect to send data from a data source to Discovery. For a list of the data sources supported by App Connect default connectors, see Connectors A-Z. For more information about integrating App Connect with Discovery, see How to use IBM App Connect with IBM Watson® Discovery.

> **Note:** To use an App Connect connector, you must create a separate App Connect instance. Costs that are incurred from a paid App Connect instance are not included with the cost of using Discovery. Except for indexing, Discovery does not support any integration with App Connect that you perform on your own.

Data source requirements

The following requirements and limitations are specific to Discovery on IBM Cloud:

- A collection can connect to only one data source.
- For more information about size limits, which can differ per plan, see the following topics:
    - Collection limits
    - Document limits

Installing IBM Secure Gateway for on-premises data

To connect to an on-premises data source, you first need to download, install, and configure IBM® Secure Gateway for IBM Cloud®.

After you install the client for one on-premises data source, you can reuse it for other data sources in the project.

The number of gateways that you can create is limited to 50.

For more information, see About Secure Gateway.

You can use the IBM Secure Gateway with the following connectors only:

- Web crawl
- Microsoft SharePoint On Prem

To install IBM® Secure Gateway for IBM Cloud®, complete the following steps:

1. From the data source configuration page, click **Manage connection**.
2. On the *Download and install Secure Gateway client* page, download the appropriate version of IBM® Secure Gateway for IBM Cloud®.
3. After you complete the download, click **Download Secure Gateway and Continue** .
4. When prompted, enter the **Gateway ID** and **Token** that are displayed.

   For more information, see Installing the client.
5. On the machine where the Secure Gateway Client is running, open the Secure Gateway dashboard at `http://localhost:9003`.
6. Click **add ACL** on the dashboard, and add the URL of the data source that you want to access to the **Allow access** list.

   For example, hostname: `mycompany.sharepoint.com` or `mycompanywebsite.com` and port: `80`.
7. Return to Discovery, and click **Continue**.
    - If the connection is successful, a `Connection successful` message is displayed.
    - If the connection is unsuccessful, open the IBM® Secure Gateway for IBM Cloud® dashboard, and verify that the endpoints on the **Allow access** list are correct.

Data source connection and data isolation

When you connect to external data sources, you reduce the data isolation of your service instance because data in transit between the source and the service cannot be isolated. All other data isolation (at-rest, administration, query) remains in full. All in-flight communication among services and data sources is encrypted with TLS v1.2. The private keys for the TLS certificates are encrypted at rest with AES-256-GCM encryption. The service certificates expire every three years and the certificate revocation lists are updated monthly. All credentials are sent over an encrypted connection that uses TLS v1.2 and are encrypted at rest with AES-256 encryption. Connections to data sources use the secure protocols that are supported by the data sources.

Viewing collections that are connected to a gateway

You can view a list of collections that are connected to a particular gateway. Complete the following steps to view collections that share a particular gateway:

1. From the **My projects** page, click **Data usage and GDPR**.
2. Click **On premises**.

   Collections that share a common gateway are displayed in the *Connected collections* list.

Connecting to data sources with IP restrictions

Some data sources allow crawlers from only a limited number of trusted network addresses or domains to access and process their data. If one of the data sources that you want to connect to limits access in this way, you can add IBM-managed IP addresses to the allowlist of the data source.

Tip: Network addresses are subject to change from time to time. You can monitor for updates to these addresses by subscribing to the repo notifications for this page. Click **Edit Topic** and then select **Watching** in the Notifications dialog of the repo.

- For service instances that are hosted in a US-based data center and that were created on or after 1 May 2020, add the following IP addresses:

```
150.238.21.0/28
169.48.255.224/28
174.36.69.128/28
```

- For service instances that are hosted in non-US data centers and that were created on or after 21 February 2021, add the following IP addresses:

```
159.122.203.64/28
158.175.114.128/28
158.176.107.48/28
```

- For a list of IP addresses that you can add to an allowlist for services instances that were created before 1 May 2020 (US) and before 21 February 2021 (non-US), see the network addresses that are listed for Cloud Foundry.
  - Refer to the **Dallas** data center IP addresses for all US-hosted service instances.
  - Refer to the **London** data center IP addresses for all service instances that are hosted outside the US.

Box

Crawl documents that are stored in a Box data source.

IBM Cloud **IBM Cloud only**

Note: This information applies only to managed deployments. For more information about connecting to Box from an installed deployment, see Box.

What documents are crawled

During the initial crawl of the content, documents from all of the folders that can be accessed from your Box application are crawled and added to your collection. Box notes are stored in JSON format, so Discovery also ingests any Box notes in the specified folders.

The following table illustrates the objects that Discovery can crawl.

Data source	Supports scheduled document refreshes?	Objects that are crawled
Box (**App access**)	No	Files, folders that you share explicitly
Box (**Enterprise access**)	Yes (New and modified documents only)	Files, folders

Table 1. Data sources crawling support

When you configure Box with App access only, you must create App Users and share the files that you want to crawl with these users. You cannot crawl Box files that are shared only by the Service Account.

For more information about access, see these Box documentation help topics:

- App Users
- Service Accounts

Documents that are deleted from Box are not deleted from the collection.

All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

Data source requirements

In addition to the data source requirements for all managed deployments, your Box data source must meet the following requirement:

You must obtain any required service licenses for the data source that you want to connect to. For more information about licenses, contact the system administrator of the data source.

Prerequisite step

You must create a custom application in Box before you can connect to Box from Discovery.

1. In Box, create a custom app that uses *Server Authentication with JWT* as its authentication method.

   For detailed steps, see Setup with JWT in the Box Developer Documentation.

Follow these guidelines when you create the app:

- During the setup procedure, choose to use the *Server Authentication with JWT* method to verify application identity with a key pair.
- When you configure the custom app, you can choose to use one of the application access levels:
  - App access only
  - App access plus Enterprise access

  Refreshing documents on a schedule is supported only when you choose **App access plus Enterprise access**.

  > ⚠ **Important:** If you set up the connection with **App access**, you must create App Users and share the files that you want to crawl with the App Users you define. With this configuration, new and modified documents are not crawled during a refresh.

  - If you are an administrator, configure **App access plus Enterprise access**. Otherwise, you can configure the app to have **App access**. However, you must get application approval from a Box administrator.
  - For both application access levels, specify the following settings:
  - Choose the following scopes:
    - *Read all folders stored in Box*
    - *Write all folders stored in Box*
    - *Manage Users*

    For apps with Enterprise access only: Add this extra scope:

    - *Manage Enterprise Properties*
  - Enable the following advanced features:
    - *Make API calls using the as-user header*
    - *Generate User Access Tokens*
- Get the custom app authorized by an administrator.

  For more information, see App approval in the Box Developer Documentation.
- After the app is created, authorized, and authentication is configured, download the app settings as a JSON file from the dev console.

  You provide the following information from this file when it is requested later:

  - `client_id`
  - `enterprise_id`
  - `client_secret`
  - `public_key_id`
  - `private_key`
  - `passphrase`

Connecting to the Box data source

From your Discovery project, complete the following steps:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **Box**, and then click **Next**.
4. Refer to the values from the Box app settings JSON file that you downloaded during the previous procedure to complete the following fields:

   Client ID

   The private key that you specify when you configure your Box app.

   Client Secret

   The client secret that you specify when you configure your Box app.

   Enterprise ID

   The enterprise ID of the Box account.

Public Key ID

The public key ID that Box generates.

Private Key

A part of the key pair that is generated to interact with the Box website.

Passphrase

The passphrase that is required to decrypt the private key if the private key is an encrypted file.

5. Click **Next**.
6. Name the collection.
7. If the language of the documents in Box is not English, select the appropriate language.

   For a list of supported languages, see Language support.
8. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.
9. Choose the folders that you want to crawl.
10. If you want to limit the types of files to add to the collection, you can list the file extensions for file types to either include or exclude.

    > ⚠ **Important:** When you choose to list extensions for file types to exclude, you must add at least one file extension.

    For a list of supported file types, see Supported file types.
11. If you want the web crawl to extract text from images on the site, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

    > **Note:** When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.
12. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Currently, not all documents are refreshed during scheduled recrawls. For more information, see the release note.

IBM Cloud Object Storage

Crawl documents that are stored in an IBM Cloud® Object Storage data source.

IBM Cloud **IBM Cloud only**

> **Note:** This information applies only to managed deployments.

What documents are crawled

During the initial crawl of the content, documents from all of the content that can be accessed from the storage endpoint are crawled and added to your collection. You cannot crawl private endpoints.

During subsequent scheduled recrawls, only new and modified documents are crawled and any changes are reflected in your collection. Documents that are deleted from the external data source are not deleted from the collection.

All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

The following table illustrates the objects that Discovery can crawl.

Data source	Objects that are crawled

IBM Cloud Object Storage	Buckets, files

Table 1. Data sources crawling support

What you need before you begin

Obtain any required service licenses for the content on the website that you want to connect to. For more information about licenses, contact the system administrator of the data source.

Endpoint

The `endpoint` for your IBM Cloud Object Storage data. For example, `s3.us-south.cloud-object-storage.appdomain.cloud`.

Do not include `http://` or `https://` in the endpoint value. For more information, see Regional Endpoints.

In addition to the endpoint, you must provide credentials to enable authentication with the object store. You can choose to use one of the following authentication methods:

HMAC

Uses a hash-based message authentication code to authenticate users. HMAC is a cryptographic authentication technique that uses a hash function and a secret key. The data is scrambled before it is sent over the internet. Then, the intended recipient uses the secret key to unscrambles the data. For more information, see HMAC authentication.

IAM

Uses the IBM Cloud Identity and Access Management (IAM) service to authenticate users. The advantage of this authentication type is that the user can use the same process to access all of the resources in the IBM Cloud Platform. For more information, see IAM authentication.

To access the credential information, go to the service credentials page of your IBM Cloud Object Storage service instance. Expand the service credential to see the credential details.

For more information, see Service credentials in the Object Storage product documentation.

HMAC authentication

If you want to use HMAC authentication, you must have the following information ready:

Access key id

The `access_key_id` that was generated when the IBM Cloud Object Storage instance was created. For example, `347aa3a4b34344f8bc7c7cccdf856e4c`.

Secret access key

The `secret_access_key` to use to sign requests. This key was generated when the IBM Cloud Object Storage instance was created. For example, `gvurfb82712ad14W7a7915h763a6i87155d30a1234364f61`.

IAM authentication

If you want to use IAM authentication, you must have the following information ready:

IAM API key

For example, `0viPHOY7LbLNa9eLftrtHPpTjoGv6hbLD1QalRXikliJ`.

Resource instance ID

For example, `cloud-object-storage:global:a/3ag0e9402tyfd5d29761c3e97696b71n:d6f74k03-6k4f-4a82-b165-697354o63903::`.

Connecting to the data source

From your Discovery project, complete the following steps:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **IBM Cloud Object Storage**, and then click **Next**.
4. Choose a credential type, and then complete the fields with the information that you collected earlier.
   - IAM
   - HMAC

   Click **Next**.
5. Name the collection.
6. If the language of the documents in storage is not English, select the appropriate language.

   For a list of supported languages, see Language support.
7. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.
8. Choose the buckets that you want to crawl.

   The more buckets that you select, the longer the processing of the documents takes.
9. If you want to limit the types of files to add to the collection, you can list the file extensions for file types to either include or exclude.

   > ⚠ **Important:** When you choose to list extensions for file types to exclude, you must add at least one file extension.

   For a list of supported file types, see Supported file types.
10. If you want the crawler to extract text from images on the site, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

    > **Note:** When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.
11. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Microsoft SharePoint Online

Crawl documents that are stored in a Microsoft SharePoint Online data source.

IBM Cloud **IBM Cloud only**

> **Note:** This information applies only to managed deployments. For more information about connecting to SharePoint Online from an installed deployment, see SharePoint Online.

What documents are crawled

During the initial crawl of the content, documents from all of the objects that can be accessed from the site collection path that you specify are crawled and added to your collection. You cannot limit the crawl to one library within a site collection, for example. All objects in the specified Site collection path are crawled. Custom metadata that is associated with the SharePoint content is crawled also. You can crawl one site collection path per collection. You cannot crawl *Personal SiteCollections*.

During subsequent scheduled recrawls, only new and modified documents are crawled and any changes are reflected in your collection. Documents that are deleted from the external data source are not deleted from the collection.

All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

Discovery can crawl the following objects:

- SiteCollections
- Sites
- SubSites
- Lists
- List Items
- Document Libraries
- List Item Attachments

Data source requirements

In addition to the data source requirements for all managed deployments, your SharePoint Online data source must meet the following requirements:

- The Site Collection that you connect to must be one that was created with an Enterprise plan. It cannot be a collection that was created with a frontline worker plan.
- You must have an Azure Active Directory user ID with permission to read all of the objects that you want to crawl. For example, `<admin_user>@.onmicrosoft.com`. The user ID does not need `SiteCollection Administrator` permission.

You can choose how to authenticate with the external Microsoft SharePoint account from the following options:

Open Authentication (OAuth v2)

Authenticates with the external data source by using a token so that your user credentials do not need to be shared. With this authentication method, you can log in to your Microsoft account directly to generate a token that is used by Discovery to connect to your data.

The *Sign in with Microsoft* option that uses Open Authentication v2 to authenticate with the external data source is a beta feature.

Before anyone can create connectors that use this authentication method, a user with the *Global Administrator* role must complete a one-time prerequisite steps to authorize the connection for all projects in the Discovery service instance.

Security Assertion Markup Language (SAML)

An older mechanism for authentication and authorization that requires user credentials to be shared with the Discovery service.

If you choose to use this authentication method, your Microsoft SharePoint account must meet the following requirements:

- Unless you created your SharePoint Online account before January 2020, two-factor authentication is enabled for the account by default. You must disable two-factor authentication.

  To view and change your multifactor authentication status, see View the status for a user or Change the status for a user.

- The crawl user account must have legacy authentication and `Contribute` level permissions enabled.

  To enable legacy authentication, go to the Azure portal or contact your SharePoint administrator.

- The connector supports the `Password hash synchronization (PHS)` method for enabling hybrid identity only. Use any other type (such as Pass-through authentication or Federation) at your own risk.

- You must know the following information:

Username

The username of the user account to use to connect to the SharePoint Online SiteCollection that you want to crawl.

For example, `<janedoe>@exampledomain.onmicrosoft.com`.

Password

The password to connect to the SharePoint Online SiteCollection that you want to crawl.

This value is never returned and is only used when credentials are created or modified.

What you need before you begin

You must have the following information ready. If you don't know it, ask your SharePoint administrator to provide the information or consult the Microsoft SharePoint developer documentation:

Organization URL

The root URL of the source that you want to crawl. Specify the domain name of the URL, for example `https://<company>.<domain>.com`.

Site collection path

The `site_collection_path` to the section of the site where you want to start the crawl.

For example, if the content that you want to crawl is available from `https://<company>.<domain>.com/sites/test`, then you can specify `https://<company>.<domain>.com` as the Organization URL and `/sites/test` as the Site collection path.

- You cannot specify folder paths as input.
- You cannot specify a path to an Active Server Page Extended (ASPX) file, such as URLs to document libraries, lists, and subsites.
- If you don't specify a path, the default value of `/` is used, and the root site collection is crawled.

- **Application ID**: ID of the data source that you want to crawl. This information is required only if you want to store ACL information that is associated with the source documents.

One-time prerequisite step for OAuth

Before anyone can configure the connector to use OAuth v2 authentication method, a user with the *Global Administrator* role in Microsoft Azure Directory where the data source is located must complete steps to register the Discovery enterprise application in Microsoft Azure. This step must be completed once per Discovery service instance.

The administrator does not need to create the application in Azure. When they choose SharePoint Online as the data source, the Discovery service generates the app automatically. As described in the procedure to follow, during the set up of the connector, the administrator must log in to Microsoft with credentials for a user with the *Global Administrator* role in Microsoft Azure Directory and allow the enterprise application to be registered.

The following steps must be completed by a global administrator one time only per service instance:

1. Review the default user access settings that will be applied to the enterprise application in Microsoft Azure.

   Enterprise applications can handle user access in many ways. Check the default settings to ensure that they are appropriate for your deployment by completing the following steps:

   1. Log in to Microsoft Azure.
   2. From the *Enterprise applications* page in *Azure Active Directory*, click *Consent and permissions*.

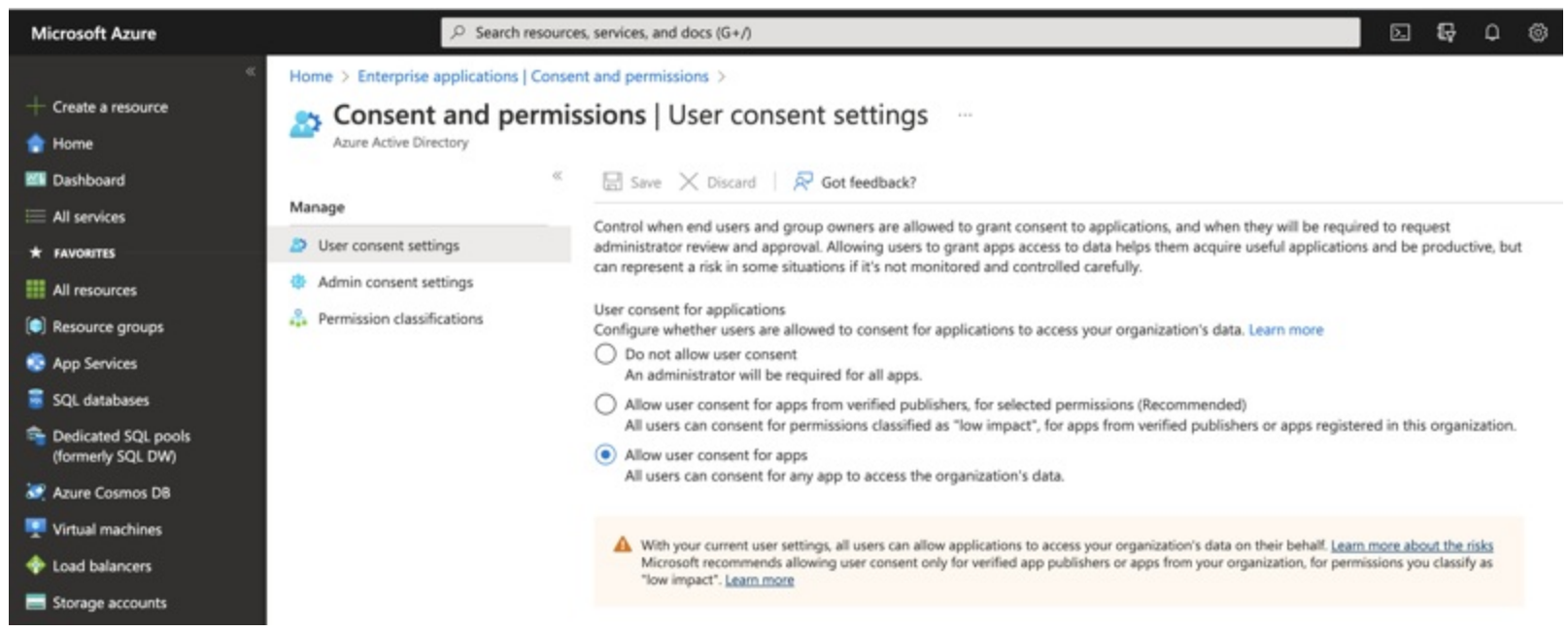

Figure 1. Microsoft Azure Enterprise application permissions user interface

1. Do one of the following things:

   - If *Allow user consent for apps* is selected, no more action is needed.
   - If *Allow user consent for apps from verified publishers, for selected permissions* is selected, then complete the following steps:

     Click *Permissions classifications* link, and then ensure that the following permissions are configured at a minimum:

     - Office 365 SharePoint Online: MyFiles.Read
     - Office 365 SharePoint Online: AllSites.Read
     - Microsoft Graph: offline_access
     - Microsoft Graph: profile

   The *Do not allow user consent option* is not supported.

The settings that you specify will be applied to the enterprise application that is created by Discovery in subsequent steps.

2. From the navigation pane of Discovery, choose **Manage collections**.
3. Click **New collection**.
4. Click **SharePoint Online**, and then click **Next**.
5. Add a URL to the **Organization URL** field.
6. Click **Sign in with Microsoft**.

   Pop-ups must be enabled for this site in your web browser.

   The *Sign in with Microsoft* option that uses Open Authentication to authenticate with the external data source is a beta feature.

   Log in to your Microsoft SharePoint account with your user name and password, and then complete two-factor authentication, if necessary.

   ⚠ **Important:** Remember, the credentials you use must have the *Global Administrator* role in Microsoft Azure Directory. If you are not prompted for a user name and password, take note. You might be logged in to a Microsoft Sharepoint account already. If you are logged in to an account that you don't want to use for this connector, stop here. (Any account where you are logged in will be used automatically. And you cannot change the account configuration later.) Open a web browser in incognito mode and start this procedure over from step 1.

   Discovery generates an enterprise application that it will register with the SharePoint organization that you specify. The enterprise application name has the format *IBM App Connect_{unique name}*.

7. Review the permissions that are associated with the enterprise application that Discovery will register, and then select **Consent on behalf of your organization**.

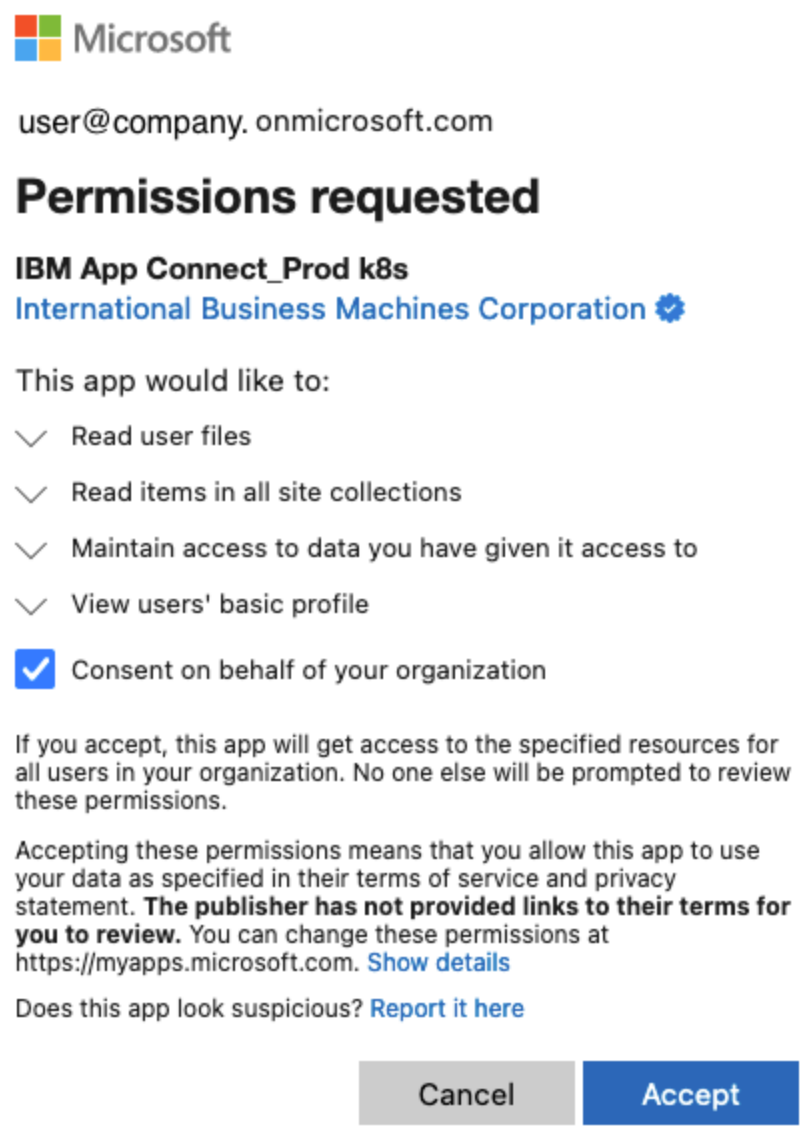

Figure 2. Discovery permission request dialog

8. Click **Accept**.
9. If you want to create a collection, you can name the collection, and then click **Finish**.

   Otherwise, you can click **Back** to exit the collection creation process.

Now, anyone from your organization who works in a project that is hosted by the same Discovery service instance can create a collection by using the SharePoint Online connector.

OAuth support revisions

Support for the OAuth method of authentication was added with a software update in February 2022. If you want to update an existing connector to use OAuth instead of SAML, you must re-create the connector. You cannot change the authentication mechanism for an existing connector.

The OAuth method of authentication was updated in January 2023. The enterprise application that is registered with Microsoft Azure now requires *Read* access only. Previously, the enterprise application required *Write* access. If you want to take advantage of this change, delete your current enterprise application and recreate the connector. For more information about how to delete an enterprise application, see the Microsoft documentation.

Connecting to the data source

To configure the Microsoft SharePoint Online data source, complete the following steps in Discovery:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **SharePoint Online**, and then click **Next**.
4. Add a URL to the **Organization URL** field.
5. To enable access to your external data source, choose the method that you want to use to authenticate with the data source from the following options:

   Open Authentication (OAuth v2)

   Click **Sign in with Microsoft**.

   Pop-ups must be enabled for this site in your web browser.

   The *Sign in with Microsoft* option that uses Open Authentication to authenticate with the external data source is a beta feature.

   Log in to your Microsoft SharePoint account with your user name and password, and then complete two-factor authentication, if necessary.

   Security Assertion Markup Language (SAML)

   Specify a username and password for a user that is authorized to access the site you want to crawl, and then click **Next**.

6. Specify the path you want to crawl in the **Site collection path** field.
7. Name the collection.
8. If the language of the documents on the site is not English, select the appropriate language.

   For a list of supported languages, see Language support.
9. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.
10. **Optional**: If you want to store any access control information that exists in the SharePoint documents that you crawl, in the *Security* section, set the **Include Access Control List** switch to `On`.

    When you enable this option, information about SharePoint access rules that is stored in SharePoint source documents is retained and stored as metadata in the documents that are added to your collection.

    This feature is not the same as enabling document-level security for the collection. The access rules in the document metadata are not used by Discovery search. Enabling this feature merely stores the information so that you can leverage the access rules when you build a custom search solution.

    ⚠ **Important:** Use of this feature increases the size of the documents that are generated in the collection and increases the crawl time. Only enable the feature if your use case requires that you store the SharePoint document ACL information.

    If you enable this feature, someone with the administrator role in Microsoft SharePoint must take extra steps to ensure that users who crawl the site have the right permissions to access ACL metadata.

    An administrator must complete the following steps:

    1. Log in to Microsoft SharePoint.
    2. Open the page for your SharePoint site.
    3. From the settings menu, choose *Site permissions*.
    4. Click *Advanced permission settings*.
    5. Make sure that people who want to collect access control information during a crawl have or are members of a group that has the *Full Control*

permission for the site.

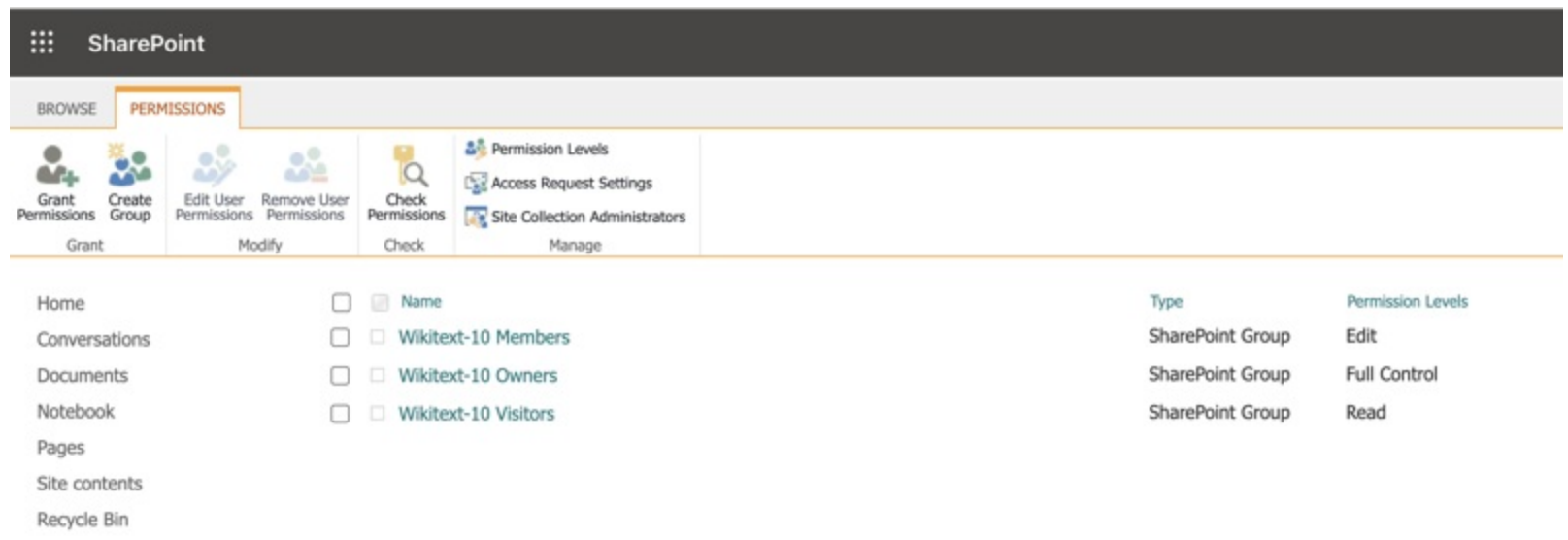

Figure 3. Microsoft SharePoint permissions user interface

Note: When access control list information is not extracted, *Read* permission is sufficient for all users who crawl the content.

11. If you want to limit the types of files to add to the collection, you can list the file extensions for file types to either include or exclude.

    Important: When you choose to list extensions for file types to exclude, you must add at least one file extension. By default, the *Extension filter* is applicable to SharePoint *Document Libraries* and *List Item Attachments* objects only. To apply the filter to all SharePoint object types, set **Apply extension filter to all SharePoint object types** to **On** on the user interface.

    For a list of supported file types, see Supported file types.

12. If you want the crawler to extract text from images on the site, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

    Note: When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.

13. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Note: You cannot currently change the user account that is associated with the OAuth setup later, nor any of the details of the existing user account that the connector is configured to use. For example, you cannot update the password that was used to set up the connection after a password change in SharePoint.

Sample access control list information

The following screen capture illustrates the type of ACL information that is stored in the document when you include the access control list.

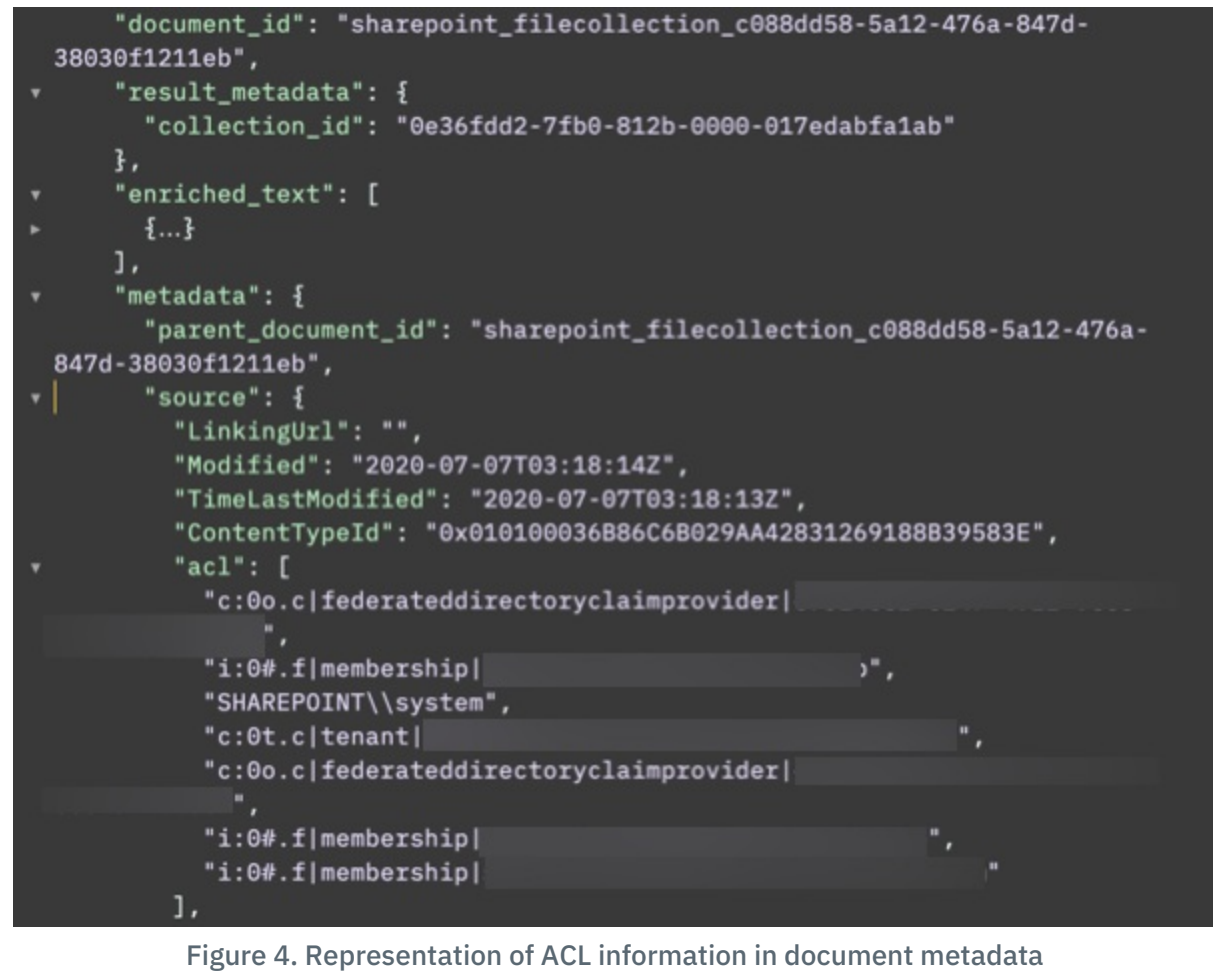

Figure 4. Representation of ACL information in document metadata

Microsoft SharePoint On Prem

Crawl documents that are stored in a Microsoft SharePoint data source that is hosted on premises.

IBM Cloud **IBM Cloud only**

> **Note:** This information applies only to managed deployments. For more information about connecting to an on-premises SharePoint data source from an installed deployment, see SharePoint On Prem.

What documents are crawled

During the initial crawl of the content, documents from all of the objects that can be accessed from the site collection path that you specify are crawled and added to your collection. Custom metadata that is associated with the SharePoint content is crawled also. You can crawl one site collection path per collection. You cannot crawl *Personal SiteCollections*.

During subsequent scheduled recrawls, only new and modified documents are crawled and any changes are reflected in your collection. Documents that are deleted from the external data source are not deleted from the collection.

All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

The following table illustrates the objects that Discovery can crawl.

Data source	Objects that are crawled
Microsoft SharePoint On Prem	SiteCollections, Sites, SubSites, Lists, List Items, Document Libraries, List Item Attachments

Table 1. Data sources crawling support

Data source requirements

In addition to the data source requirements for all managed deployments, your SharePoint On Prem data source must meet the following requirements:

- You can connect to a SharePoint 2013, 2016, or 2019 on-premises data source.
- The user ID must have `SiteCollection Administrator` permission and be able to access all of the sites and lists that they want to crawl.
- The crawler supports Windows New Technology LAN Manager (NTLM) v1 authentication only. It does not support NTLM v2 or Security Assertion Markup Language (SAML) authentication.

What you need before you begin

You must have the following information ready. If you don't know it, ask your SharePoint administrator to provide the information or consult the Microsoft SharePoint developer documentation:

Username

The username to use to connect to the SharePoint On Prem web application that you want to crawl. For example, `siteadmin01`.

Password

The password to connect to the SharePoint On Prem web application that you want to crawl. This value is never returned and is only used when credentials are created or modified.

Web Application URL

The SharePoint web application URL. For example, `https://sharepointwebapp.com:8443`. If you do not enter a port number, the default value of `80` is used for an HTTP URL and `443` for HTTPS.

Domain

The domain name of the SharePoint On Prem account. For example, `sharepoint.mycointernal`.

Prerequisite step

Before you can connect to a SharePoint On Prem data source, you must install and configure IBM® Secure Gateway for IBM Cloud®.

For more information, see Installing IBM Secure Gateway for on-premises data.

Connecting to the data source

To configure the Microsoft SharePoint On Prem data source, complete the following steps in Discovery:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **SharePoint On Prem**, and then click **Next**.
4. Add values to the following fields:
   - Username
   - Password
   - Web Application URL
   - Domain

   Click **Next**.
5. Name the collection.
6. If the language of the documents on the site is not English, select the appropriate language.

   For a list of supported languages, see Language support.
7. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.
8. If you want to limit the types of files to add to the collection, you can list the file extensions for file types to either include or exclude.

   Important: When you choose to list extensions for file types to exclude, you must add at least one file extension.

   For a list of supported file types, see Supported file types.
9. If you want the crawler to extract text from images on the site, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

   Note: When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.
10. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Salesforce

Crawl documents that are stored in a Salesforce data source.

IBM Cloud **IBM Cloud only**

> **Note:** This information applies only to managed deployments. For more information about connecting to Salesforce from an installed deployment, see Salesforce.

What documents are crawled

During the initial crawl of the content, documents from all of the objects that can be accessed from the URL that you specify are crawled and added to your collection. Knowledge Articles are crawled only if their **version** is `published` and their languages is `en-us`.

During subsequent scheduled recrawls, only new and modified documents are crawled and any changes are reflected in your collection. Documents that are deleted from the external data source are not deleted from the collection.

All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

Discovery can crawl the following objects:

- Any default and custom objects that you have access to
- Accounts
- Contacts
- Cases
- Contracts
- Knowledge articles
- Attachments

Data source requirements

In addition to the data source requirements for all managed deployments, your Salesforce data source must meet the following requirements:

- The instance that you plan to connect to must be part of an Enterprise plan or higher.
- You must obtain any required service licenses for the data source that you want to connect to. For more information about licenses, contact the system administrator of the data source.

What you need before you begin

You must have the following information ready. If you don't know it, ask your Salesforce administrator to provide the information or consult the Salesforce developer documentation.

Username

: The `username` of an account that has access to the Salesforce site. For example, `jdoe@example.com`

Password

: The password associated with the username. For example, `myP@ssw0rd`.

Service token

: A valid Salesforce security token. For example, `mnaO8jsRET5CiJww9JnURlNN`.

URL

: The URL of the Salesforce site that you want to crawl. For example, `https://my.salesforce.com`

Connecting to the data source

To configure the Salesforce data source, complete the following steps in Discovery:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **Salesforce**, and then click **Next**.
4. Add values to the following fields:
    - Username
    - Password plus service token

      To form the password, concatenate the Password and Service token values that you noted earlier. For example,

`myP@ssw0rdmnaO8jsRET5CiJww9JnURlNN`. The password and token values are never returned and are used only when credentials are created or modified.

- URL

Click **Next**.

5. Name the collection.
6. If the language of the documents on the site is not English, select the appropriate language.

   For a list of supported languages, see Language support.
7. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.
8. Select the objects that you want to crawl.

   The more objects that you select, the longer the processing of the documents takes.
9. If you want to limit the types of files to add to the collection, you can list the file extensions for file types to either include or exclude.

   ⚠ **Important:** When you choose to list extensions for file types to exclude, you must add at least one file extension.

   For a list of supported file types, see Supported file types.
10. If you want the crawler to extract text from images on the site, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

    Note: When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.
11. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Web crawl

Add a web crawl collection to crawl a website, analyze its page content, and store meaningful information. Specify one or more base web page URLs and configure how many linked pages for the web crawl to follow. You can configure how often to synchronize with the website, so you control how up to date the data in your collection is.

IBM Cloud **IBM Cloud only**

Note: This information applies only to managed deployments. For more information about connecting to a website from an installed deployment, see Web crawl.

What documents are crawled

You can connect to the following types of web content:

- Public websites
- Private company websites or other sites that require authentication
- Websites that are behind a corporate firewall

During the initial crawl of the content, all website pages that match your search settings are crawled and added to the document index of your collection. The crawl starts on the web page that you specify in the *Starting URLs* field. If your collection is configured to follow links, the crawl follows links on the starting page that share the same subtree as the starting page. For example, if you specify `https://www.example.com/banking/faqs.html`, links with URLs that begin with `https://www.example.com/banking/` are crawled. If you specify `https://www.example.com/banking`, links with URLs that begin with `https://www.example.com/` are crawled.

The crawl cannot access secure subdirectories. For example, if a subdirectory that you expect the crawl to access, such as `https://www.example.com/banking/pdfs`, isn't being crawled, check whether you can access the subdirectory URL from a web browser directly. If you can't access it, the crawl can't access it.

During subsequent scheduled recrawls, a full recrawl is performed and any changes are reflected in your collection. Documents that were added to your collection from website pages that are later deleted from the external website are not deleted from the collection. However, starting with collections that were created after April 2022, when you remove a starting URL from the web crawl configuration, any associated documents are deleted. Deleted

documents include indexed documents that were added to the collection based on the content of the web page at the starting URL and documents that were derived from web pages that the starting URL linked to. You cannot limit the number of indexed documents by changing other settings, such as changing the existing URL to include a path with a more limited scope than before or reducing the maximum number of links to follow to 0. Only by deleting the URL can you remove the indexed documents that are associated with it.

The web crawler can crawl web pages that use JavaScript to render content, but the crawler works best on individual pages, not entire websites. It cannot crawl sites that use dynamic URLs; if you can't see any content when you view the source code of a web page in your browser, then the service cannot crawl it.

If you want to crawl a group of URLs that includes some websites that require authentication and some that don't, consider creating a different collection for each authentication type. The connector does not support cookie-based crawling.

All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

The following table illustrates the objects that Discovery can crawl.

Objects that are crawled
Websites, website subdirectories

Table 1. Data sources crawling support

Prerequisite step

If you want to connect to a website that is hosted behind a firewall, set up an IBM® Secure Gateway for IBM Cloud® connection first.

Valuable content is often stored on your company's internal website. Typically, such intranet websites are accessible only from a computer that is connected to your office network or through a VPN connection. You can establish a persistent and more secure connection between the web crawler and this type of internal site by using Secure Gateway.

For more information about how to set up the connection, see Installing IBM Secure Gateway for on-premises data.

Connecting to the data source

To configure the web crawl collection, complete the following steps:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **Web crawl**, and then click **Next**.
4. Name the collection.
5. If the language of the content on the website is not English, select the appropriate language.

   For a list of supported languages, see Language support.
6. **Optional**: You can change the synchronization schedule.

   For more information, see Crawl schedule options.
7. Specify the URL of the website that you want to crawl.
   - If the site you want to crawl requires a login, set **Basic authentication** to `On`, add the URL of the page to the **Starting URL** field, and then click **Add**.

     Add a username and password with access to the site, and then click **Save credentials**. You can specify only one set of credentials per collection.

     For example, you can specify `https://cloud.ibm.com` as the starting URL and add your IBMid as the credentials.

     If you want to start the crawl from a specific section of the site, specify it in the **Starting URLs** field. The domain name of the subsection must match the domain in the URL you specified earlier.

     For example, you might change the starting URL to `https://cloud.ibm.com/unifiedsupport/supportcenter`.
   - For any public web pages that you want to crawl, add the URL for the root page of the website to the **Starting URLs** field, and then click **Add**. You can add more than one starting page.

     The final forward slash (`/`) in the URL determines the subtree to crawl. If you specify `https://www.example.com/banking/faqs.html`, all URLs that begin with `https://www.example.com/banking/` are crawled, for example. If you specify `https://www.example.com/banking` all URLs that begin with `https://www.example.com/` are crawled.

By default, the number of consecutive links that the crawl follows from the starting URL is `2`. To change the number of hops or to list website sections to exclude from the crawl, click the edit icon.

- The maximum number of hops allowed is `20`.
- To specify URL paths to exclude, add the site path. For example, if the starting URL is `https://example.com`, you can exclude `https://example.com/pricing` by entering `/pricing/`.

  Any section of the web address that contains the site path you specify is excluded. For example, if you specify `/licenses/`, the page `https://example.com/products/licenses/europe` is excluded, among others.
- If you want to restrict the crawl to a single page, add the URL to the **Starting URLs** field. For example, `https://www.example.com/banking/faqs.html`. Click the edit icon to set the **Maximum number of links to follow** to `0`.

- If the website that you want to crawl uses JavaScript to customize the page content before it is displayed, you must take an extra step.

  After you enter the starting URL and click **Add**, edit the URL by clicking the edit icon. Set the *Execute JavaScript during crawl* switcher to **On**, and then click **Save**.

  Note: When JavaScript processing is enabled, it takes 3 to 4 times longer to crawl a page. Use it only on individual web pages where you know it is necessary because the page renders its content dynamically. If you see timeout messages or the crawl ends without adding content to the collection, decrease the number of web pages that are included in the crawl. For example, you can specify the exact page to crawl in the *Starting URLs* field, and set *Maximum number of links to follow* to 0.

- To connect to a website that is hosted behind a firewall, set up an IBM® Secure Gateway for IBM Cloud® connection first.

  Expand *More connection settings*, and then set **Connect to on-premises network** to `On`. Provide details about your Secure Gateway connection.

8. Optional: Add another web address to the **Starting URLs** field.

   Important: The number of starting URLs for a single collection must be less than 100. If you have a requirement to crawl a large number of websites, see I need to crawl lots of sites. What's my limit?.

   The number of web pages that are crawled is limited to 250,000, so the web crawler might not crawl all the specified websites.

   The number of child URLs per URL that are crawled is limited to 10,000. If the number of child URLs within any crawled URL exceeds 10,000, the crawler cannot process any of the content in the child URLs.

9. If you want to limit the types of files to add to the collection, you can list the file extensions for file types to either include or exclude.

   Important: If the URLs for your website pages do not end in *.html*, use the exclude filter instead of the include filter. You must add at least one file extension to exclude.

   For a list of supported file types, see Supported file types.

10. If you want the web crawl to extract text from images on the site, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

    Note: When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.

11. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

I need to crawl lots of sites. What's my limit?

The service can support a total of 500 crawler connections per Discovery service instance. All of the data sources except Web crawl use one crawler connection each. For Web crawl, one connection is required for every 5 starting URLs. If you add 10 starting URLS, for example, Discovery generates the extra crawler connection that is needed to support the extra 5 URLs. Therefore, the maximum number of starting URLs that you can use depends on the other data collections that are configured in your service instance. You can calculate the limit yourself.

To calculate the starting URL limit, complete the following steps:

1. Calculate the number of other data source collections in the service instance, meaning this project and any other projects in the same Discovery instance.

For example, you might have 2 IBM Cloud Object Store collections in one project and 2 Salesforce collections and 1 SharePoint Online collection in another project. In this example, the total number of other data source collections is 5.

2. Subtract the number of other data source collections from the maximum allowed number of crawler connections, which is 500.

   For example, 500 - 5 = 495.

3. Multiply the remainder by 5 to determine the total number of starting URLs that you can use.

   For example, 495 x 5 = 2,475.

Note: To use the maximum-allowed number of starting URLs in the example, you would need 25 web crawl collections because each collection allows a maximum of 100 starting URLs to be configured. However, don't configure your instance to use the absolute maximum number allowed. If one or more additional data sources are added subsequently to a project in this service instance, it will impact the number of starting URLs that the instance can crawl successfully.

Troubleshooting crawler issues

A 403 Forbidden error is returned

The website that you want to crawl might block requests from all but a specific set of named entities. If possible, add the crawler to the allowlist for the site. The identifying header for the crawler is `User-Agent: IBM-AppConnect/V1`.

Configuring IBM Cloud Pak for Data data sources

Overview of Cloud Pak for Data data sources

In Discovery for Cloud Pak for Data, you can crawl documents from a local source that you upload or from a remote data source that you connect to. Learn more about the supported data sources and how to configure them.

IBM Cloud Pak for Data **IBM Cloud Pak for Data only**

Note: This information applies only to installed deployments. For more information about IBM Cloud data sources, see Overview of the IBM Cloud data sources.

Important: All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

You can use Discovery for Cloud Pak for Data to crawl from the following data sources:

- Box
- Database
- FileNet P8
- LDAP directory
- Local File System
- Notes
- Salesforce
- SharePoint Online
- SharePoint On Prem
- Web crawl
- Windows File System

Your data source isn't listed? You can work with a developer to create a custom connector. For more information, see Building a Cloud Pak for Data custom connector.

If you have special requirements when you add source documents, such as a need to exclude certain files, you can work with a developer to create a custom crawler plug-in. The crawler plug-in can apply more nuanced rules to what documents and what fields in the documents get added. For more information, see Building a Cloud Pak for Data custom crawler plug-in .

Data source requirements

The following requirements and limitations are specific to IBM Watson® Discovery:

- The individual file size limit is 32 MB per file, which includes compressed archive files (ZIP, CZIP, TAR). When decompressed, the individual files within compressed files cannot exceed 32 MB per file. This limit is the same for collections in which you upload your own data.

- Depending on the type of installation (starter or production mode), the number of collections you can ingest simultaneously varies. A starter installation includes one `crawler` pod, which allows three collections to be processed simultaneously. A production installation includes two `crawler` pods, which can process six collections simultaneously.

  If you are running a starter installation and you want to process more than three collections simultaneously, you must increase the number of `crawler` pods by running the following commands:

  ```
  $ oc patch wd wd --type=merge --patch='{"spec": {"ingestion": {"crawler": {"replicas": <number-of-replicas> } } } }'
  ```

  > **Note:** In a starter installation, the maximum number of simultaneous collections that can crawl an external data source is 3. If you start a fourth, that collection does not start to process until the prior three crawls finish.

  Each `number-of-replicas` allows 3 simultaneous crawls, so `number-of-replicas=2` increases the replicas to 6, and `number-of-replicas=3` increases them to 9.

Crawler plug-in settings

When you deploy one or more crawler plug-ins, you can configure your collection to use one of the plug-ins.

These settings are only available when crawler plug-ins are deployed.

- For more information about building a plug-in, see Building a Cloud Pak for Data crawler plug-in.
- For more information about deploying a crawler plug-in, see Commands and options for managing your crawler plug-ins.

When you are ready to configure a collection to use a crawler plug-in that was created by using the `scripts/manage_crawler_plugin.sh` script, you can see a *Plug-in settings* section with the following options:

- **Enable plug-in**: The switch is set to **Off**. Enable this option if you want to use a crawler plug-in to process documents.
- **Plug-in**: Lists the names of available crawler plug-ins. Select a plug-in to use.

Supporting document-level security

If document-level security is activated, you can use the security settings from your source documents to control the search results that are returned to different users.

Discovery supports prefiltering only. To prefilter, Discovery replicates the document's source access control list (ACL) at crawl time into the index. The search engine must compare user credentials to the replicated document ACLs. Discovery is faster when documents are prefiltered and when you control which documents you add to the index. However, it is difficult to model all of the security policies of the various data sources in the index and implement comparison logic uniformly. Also, prefiltering is not as responsive to changes that occur in the source ACLs after the most recent crawl.

Document-level security is supported by the following data source types:

- Box
- FileNet P8
- HCL Notes
- Microsoft SharePoint Online
- Microsoft SharePoint On Prem
- Microsoft Windows File System

> **Important:** When you query collections where document-level security is enabled, no results are returned if the users associated with your Discovery instance are not present in the source system. For more information about querying these collections, see Querying with document-level security enabled.

To enable document-level security, you must complete the following steps:

1. Create Discovery users that match the users available on the source system.
2. Associate users with your Discovery instance. For more information, see Giving users access to a Watson Discovery instance.
3. Enable document-level security for the data source when you connect to it.

Creating users for document-level security

You must create users that match the users available on the source system that Discovery is connecting to so that they can query with document-level security enabled.

1. Log in to Discovery as an administrator.
2. Create users who match the users available on your source or who are connected to the identity provider that your source system uses. If you create users for document-level security, keep the following points in mind:
   - Optional: For each user that you want to have access to query results, you must add users. The username must match the username that the source uses. This option is only for development and testing purposes. To create users individually, see Managing users.

- To connect to an identity provider that the source is using, see Connecting to your identity provider.

Note: Discovery does not synchronize changes that are made to the users in the identity provider with the user list for the service. Discovery administrators must ensure that the user list is current and remove any noncurrent users.

Box

Crawl documents that are stored in a Box data source.

IBM Cloud **IBM Cloud only**

Note: This information applies only to managed deployments. For more information about connecting to Box from an installed deployment, see Box.

What documents are crawled

During the initial crawl of the content, documents from all of the folders that can be accessed from your Box application are crawled and added to your collection. Box notes are stored in JSON format, so Discovery also ingests any Box notes in the specified folders.

The following table illustrates the objects that Discovery can crawl.

Data source	Supports scheduled document refreshes?	Objects that are crawled
Box (**App access**)	No	Files, folders that you share explicitly
Box (**Enterprise access**)	Yes (New and modified documents only)	Files, folders

Table 1. Data sources crawling support

When you configure Box with App access only, you must create App Users and share the files that you want to crawl with these users. You cannot crawl Box files that are shared only by the Service Account.

For more information about access, see these Box documentation help topics:

- App Users
- Service Accounts

Documents that are deleted from Box are not deleted from the collection.

All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

Data source requirements

In addition to the data source requirements for all managed deployments, your Box data source must meet the following requirement:

You must obtain any required service licenses for the data source that you want to connect to. For more information about licenses, contact the system administrator of the data source.

Prerequisite step

You must create a custom application in Box before you can connect to Box from Discovery.

1. In Box, create a custom app that uses *Server Authentication with JWT* as its authentication method.

   For detailed steps, see Setup with JWT in the Box Developer Documentation.

Follow these guidelines when you create the app:

- During the setup procedure, choose to use the *Server Authentication with JWT* method to verify application identity with a key pair.
- When you configure the custom app, you can choose to use one of the application access levels:
  - App access only
  - App access plus Enterprise access

  Refreshing documents on a schedule is supported only when you choose **App access plus Enterprise access**.

  Important: If you set up the connection with **App access**, you must create App Users and share the files that you want to crawl with the App Users you define. With this configuration, new and modified documents are not crawled during a refresh.

    - If you are an administrator, configure **App access plus Enterprise access**. Otherwise, you can configure the app to have **App access**. However, you must get application approval from a Box administrator.
    - For both application access levels, specify the following settings:
    - Choose the following scopes:
        - *Read all folders stored in Box*
        - *Write all folders stored in Box*
        - *Manage Users*

      For apps with Enterprise access only: Add this extra scope:

        - *Manage Enterprise Properties*
    - Enable the following advanced features:
        - *Make API calls using the as-user header*
        - *Generate User Access Tokens*
- Get the custom app authorized by an administrator.

  For more information, see App approval in the Box Developer Documentation.
- After the app is created, authorized, and authentication is configured, download the app settings as a JSON file from the dev console.

  You provide the following information from this file when it is requested later:

    - `client_id`
    - `enterprise_id`
    - `client_secret`
    - `public_key_id`
    - `private_key`
    - `passphrase`

Connecting to the Box data source

From your Discovery project, complete the following steps:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **Box**, and then click **Next**.
4. Refer to the values from the Box app settings JSON file that you downloaded during the previous procedure to complete the following fields:

   Client ID

   The private key that you specify when you configure your Box app.

   Client Secret

   The client secret that you specify when you configure your Box app.

   Enterprise ID

   The enterprise ID of the Box account.

   Public Key ID

   The public key ID that Box generates.

   Private Key

   A part of the key pair that is generated to interact with the Box website.

   Passphrase

   The passphrase that is required to decrypt the private key if the private key is an encrypted file.

5. Click **Next**.
6. Name the collection.

7. If the language of the documents in Box is not English, select the appropriate language.

   For a list of supported languages, see Language support.

8. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.

9. Choose the folders that you want to crawl.

10. If you want to limit the types of files to add to the collection, you can list the file extensions for file types to either include or exclude.

    > ⚠ **Important:** When you choose to list extensions for file types to exclude, you must add at least one file extension.

    For a list of supported file types, see Supported file types.

11. If you want the web crawl to extract text from images on the site, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

    > **Note:** When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.

12. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Currently, not all documents are refreshed during scheduled recrawls. For more information, see the release note.

Database

Crawl documents that are stored in a database that supports the Java Database Connectivity (JDBC) API.

IBM Cloud Pak for Data **IBM Cloud Pak for Data only**

> **Note:** This information applies only to installed deployments.

What documents are crawled

- Each row in the database is crawled and added to the collection as one document. The columns are indexed as metadata.
- The crawler attempts to crawl and index content, such as BLOB/BINARY, that is stored in the database. File types that are supported by Discovery are indexed. For more information, see Supported file types.
- When a source is recrawled, new documents are added, updated documents are modified to the current version, and deleted documents are deleted from the collection's index.
- All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

Data source requirements

In addition to the data source requirements for all installed deployments, your database data source must meet the following requirements:

- Discovery supports the following data source versions:
  - Data Virtualization on IBM Cloud Pak for Data 1.8.0, 1.8.3 which use Db2 11.5
  - IBM Db2: 10.5, 11.1, 11.5
  - Microsoft SQL Server: 2012, 2014, 2016, 2017
  - Oracle Database: 12c, 18c, 19c
  - PostgreSQL: 9.6, 10, 11

  > **Note:** Support for Data Virtualization was added with IBM Cloud Pak for Data 4.5.x releases

- You must obtain any required service licenses for the data source that you want to connect to. For more information about licenses, contact the system administrator of the data source.

Prerequisite step

- Decide which database tables you want to crawl. You can crawl multiple tables in a collection, and you can specify tables that have different schemas, or sets of columns. You must know the following information:

- Schema names
- Table names

For Data Virtualization on IBM Cloud Pak for Data, you can get these details from the IBM Cloud Pak for Data web client. Click the main menu icon, expand Data, and then select *Data virtualization*. At the start of the page, choose to show *Virtualized data*.

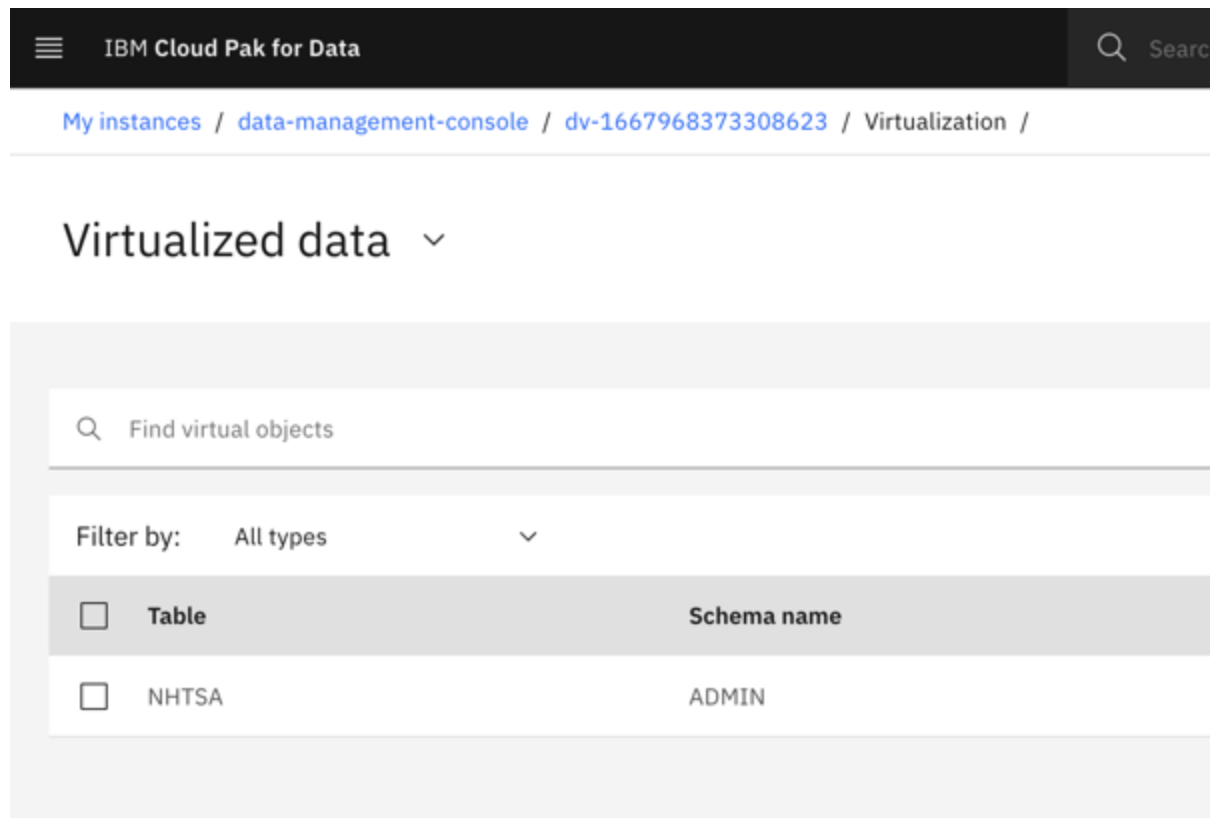

Figure 1. Virtualized data view in Cloud Pak for Data

- Be careful if you plan to crawl multiple tables that have columns with the same name but different data types. In Content Mining projects, columns with the same name but different data types are assigned to fields that have a data type suffix in the name, such as `DATA_string`. In all other project types, the data in one of the tables is excluded from the index. For example, if you have two tables that have columns that are called `DATA` and the `DATA` column in one table is populated with dates and the column in the other table is populated with strings, the data in one of the tables is excluded from the index.
- Get the user credentials for a user who has permission to access the tables that you want to crawl.
- Before you can connect to a database, you must get the JDBC driver library for the database. When you set up the database data source, you are asked to specify the JDBC driver class path.
- Before you can connect to the Data Virtualization service by using JDBC, you must install IBM Data Server driver packages. For more information, see Connecting applications to the Data Virtualization service.
- If you want to connect to an instance of Data Virtualization that is hosted in a different cluster from your Discovery service, you must forward traffic that is routed for Data Virtualization from an external infrastructure node to the master nodes of your cluster. For more information, see Updating HAProxy configuration file.

1. Download the JAR files for the JDBC driver library from the database server or vendor's website.

   The following files are associated with each database:

   - Db2 and Data Virtualization: `db2jcc4.jar`
   - Oracle: `ojdbc8.jar`
   - SQL Server: `mssql-jdbc-7.2.2.jre8.jar`
   - PostgreSQL: `postgresql-42.2.6.jar`

2. Compress the JAR files into a single compressed file.

   If you have a JDBC driver that has only one JAR file, skip this step.

3. Make a note of where the driver is stored. You must specify the directory where you store this JAR or compressed file in the next procedure so that Discovery can upload it.

Connecting to a database data source

⚠ **Important:** Before you begin, if you plan to apply enrichments to your data, create the collection in a Content Mining project type. If you are using a different project type and plan to apply enrichments, stop here. For more information, see Applying enrichments to content from a database.

From your Discovery project, complete the following steps:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.

3. Click **Database**, and then click **Next**.
4. Name the collection.
5. If the language of the documents in the database is not English, select the appropriate language.

   For a list of supported languages, see Language support.
6. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.
7. Complete the following fields in the *Enter your credentials* section:

   Database URL

   The URL of the database server.

   The following table shows example database URLs:

Database	Syntax	Example
Data virtualization (same cluster)	`jdbc:db2://{fully-qualified-hostname-of-dv-service}:{jdbc-nonssl-internal-port}/bigsql`	`jdbc:db2://c-db2u-dv-db2u-engn-svc.myproject.svc.cluster.local:50000/bigsql`
Data virtualization (separate cluster)	jdbc:db2://{cluster-address}:{jdbc-nonssl-external-port}/bigsql	jdbc:db2://api.conn.cp.example.com:30269/bigsql
Db2	`jdbc:db2://{server}:{port}/{database_name}`	`jdbc:db2://localhost:50000/sample`
Oracle	`jdbc:oracle:thin:@//{host}:{TCPport}/{service_name}`	`jdbc:oracle:thin:@localhost:1521/sample`
SQL Server	`jdbc:sqlserver://{serverName}[{instanceName}]:{port}[;property=value]`	`jdbc:sqlserver://localhost:1433;DatabaseName=sample`
Postgresql	`jdbc:postgresql://{host}:{port}/{database}`	`jdbc:postgresql://localhost/sample`

Example database URLs

   User

   The username that you obtain from the database you selected. You use this username to crawl the source. Your username is different from database to database.

   Password

   The password that is associated with your username. Your password is different from database to database.

8. Complete the following fields in the *Connection settings* section:

   JDBC driver type

   Choose the database.

   Db2 is selected by default. If you want to crawl from a database type that is not listed, select **OTHER**. To crawl data that is managed by Data Virtualization on IBM Cloud Pak for Data, keep `Db2` selected.

   JDBC driver classname

   The JDBC driver class name that is associated with the database you selected. This field is autofilled, unless you select **OTHER**.

JDBC driver classpath

Upload a JDBC driver file, which can have a .jar or .zip file extension. Alternatively, you can reuse a .jar or .zip file that you uploaded previously.

9. Complete the following fields in the *Specify what you want to crawl* section, and then click **Add**:

Schema Name

The schema that you want to crawl.

Table Name

The table within a schema that you want to crawl.

Click the edit icon to specify more table crawl settings, including:

Primary key

The primary key of the target database table. If the primary key is not configured in the target database table, you must specify the key in this field. The JDBC database crawler appends this primary key value to the URL of each crawled row to keep its uniqueness. When the primary key is a composite key, concatenate the key names by using a comma, for example `key1,key2`. If unspecified, the project defaults to the primary key fields of the table. If the primary key is configured in the target database table, this key is automatically detected.

Row filter

Optional. Specify the `SQL WHERE` clause to designate which table rows to crawl. You must specify a Boolean expression that can be the condition of a `WHERE` clause in a `SELECT` statement. If there is an error in syntax or column names, the table is excluded from the crawl, and no documents are indexed.

Column with data to extract

Name of the column with data that you want to crawl. If you don't specify the column, a column with text or with a single large object is chosen to be crawled.

MIME type of data

Optional. The MIME type is detected if not specified.

The values that you specify in the table crawl settings dialog are not displayed with the schema and tables names, but the values are applied to the database connection.

Note: The *Column with data to extract* and *MIME type of data* fields were added with the 4.6.5 release.

10. If you want the crawler to extract text from images in documents, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

Note: When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.

11. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Using Windows Authentication on Linux

The JDBC driver from Microsoft does not support Windows Authentication on Linux. If you want to use Microsoft Windows authentication to access your SQL Server on Linux, you can use a third-party JDBC driver called jTDS from Sourceforge. Specify the following values during the configuration:

- Database URL: `jdbc:jtds:sqlserver://<host>:<port>;databaseName=<database>;domain=<domain>;useNTLMv2=true;`
- JDBC driver type: `OTHER`
- JDBC driver class name: `net.sourceforge.jtds.jdbc.Driver`

Applying enrichments to content from a database

If you use a database as your data source and want to apply enrichments to the nested fields that are indexed from the database, you must use a Content Mining project type.

If your goal is to create a search application by using a Document Retrieval project type, create a Content Mining project type first. From the Content Mining project, you can connect to the database and enrich the data. Then, you can reuse the enriched collection from a Document Retrieval project.

To enrich database content for use in a Document Retrieval project, complete the following steps:

1. Create a Content Mining project.

   For more information, see Creating a project.

2. Connect to a database data source.

   For more information, see Configuring a data source: Database.

3. Apply enrichments.

   For more information, see the following topics:

   - Adding domain-specific resources
   - Applying prebuilt enrichments.

4. Create a Document Retrieval project.

   For more information, see Creating a project.

   > **Note:** When you are prompted to choose a collection, choose **Reuse data from an existing collection** . If necessary, scroll to see this option.

5. Select the collection that you created and enriched by using the Content Mining project, and then click **Finish**.

FileNet P8

Crawl documents that are stored in FileNet P8.

IBM Cloud Pak for Data **IBM Cloud Pak for Data only**

> **Note:** This information applies only to installed deployments.

What documents are crawled

- Only file types that are supported by Discovery are crawled; all others are ignored. For more information, see Supported file types.
- Document-level security is supported. When this option is enabled, your users can crawl and query the same content that they can access when they are logged in to FileNet. Discovery does not support role-based security when you crawl FileNet P8.

  For more information about document-level security, see Supporting document-level security.
- Only files with file extensions that match the file extension filter rules that you specify are crawled. *Added with the 4.7.0 release.*
- When a source is recrawled, new documents are added, updated documents are modified to the current version, and deleted documents are deleted from the collection's index.
- All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

Data source requirements

In addition to the data source requirements for all installed deployments, your FileNet P8 data source must meet the following requirements:

- The data source can crawl FileNet P8 5.5.0 and the Content Engine Web Services (CEWS) of a FileNet server that is installed on IBM Cloud Pak for Automation.
- FileNet P8 5.5.0 and FileNet on Cloud Pak for Automation support the HTTP and HTTPS protocols.

Prerequisite steps

If you want to enable document-level security, you must take some steps to set it up. For more information, see About document-level security.

Connecting to a FileNet P8 data source

From your Discovery project, complete the following steps:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **FileNet P8**, and then click **Next**.
4. Name the collection.
5. If the language of the documents in FileNet is not English, select the appropriate language.

   For a list of supported languages, see Language support.
6. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.
7. Complete the following fields in the *Enter your credentials* section:

   Content Engine Web Service URL

   The Content Engine web service URL of the IBM FileNet P8 server.

   When you enter the URL, use the format: `<protocol>://<server>:<port>/wsi/FNCEWS40MTOM`. You can use the HTTP or HTTPS protocol. The `<server>` is the hostname of the server where the Content Platform Engine is deployed and the `<port>` is the HTTP port that the application server uses, or where the Content Platform Engine is deployed.

   User

   The username to use to crawl the FileNet P8 server. You can obtain your username from your FileNet administrator.

   Password

   The password that is associated with the user.

8. In the *Specify what you want to crawl* section, enter the display name of the object store that you want to use to create, search, retrieve, and store documents in the **ObjectStore Name** field.
9. In **Crawler Space Type**, select either **Folder** or **Class**.
10. Complete the following field:

    Folder subpath or Subclass name

    The subfolder path that you can specify under RootFolder that crawls all documents that belong to the specified folder or the custom subclass of the `Document` class that crawls all documents that belong to the specified class. Before you specify anything in this field, keep in mind the following items:

    - You can specify multiple crawler spaces by using both the `Class` and `Folder` types and crawl the documents belonging to the folder name and class name.
    - You cannot specify a class outside the object store that you defined.
    - No support is available for specifying a class that is a subclass of a `Custom Object` and `Folder`.
11. After you enter one or more paths, click **Add**.
12. **Optional**: In the *Security* section, if you want to enable document-level security, set the **Enable Document Level Security** switch to `On`.

    When set to **On**, your users can crawl the same content that they have access to in FileNet.
13. If you want to limit the types of files to add to the collection, you can list the file extensions for file types to either include or exclude.

    For a list of supported file types, see Supported file types.

    > **Note:** Support for this option was added with the 4.7.0 release.
14. If you want the crawler to extract text from images in documents, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

> **Note:** When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.

15. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

LDAP directory

Crawl records in an external directory that supports the Lightweight Directory Access Protocol (LDAP).

IBM Cloud Pak for Data **IBM Cloud Pak for Data only**

> **Note:** This information applies only to installed deployments.

As the directory data is added to your collection, Discovery interprets and stores key attributes of each record according to the configuration that you specify. Later, you can find relevant records by filtering on the attributes that are of interest to you. For example, you can capture department and location information, and then filter records by location later.

For more information about the Lightweight Directory Access Protocol, see RFC 4511.

What documents are crawled

- Each LDAP record is crawled and added to the collection as one document.
- All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

Data source requirements

In addition to the data source requirements for all installed deployments, your Salesforce data source must meet the following requirements:

- The LDAP directory data source supports connections to the following types of directories:
  - IBM Security Directory Server
  - Microsoft Active Directory (On premises only)
  - Oracle Directory Server
- The LDAP directory data source collection does *not* support the following capabilities:
  - Document-level security
  - Mutual authentication. Verifying the server certificate is supported, but also verifying the client certificate is not.
  - Proxy server access to the data source

Prerequisite step

When you set up the collection, you must provide details such as the LDAP host name and port, for your directory server type. For more information about how to discover these values, see the documentation from the vendor:

- IBM Security Directory Server
- Microsoft Active Directory
- Oracle Directory Server

Connecting to an LDAP directory data source

From your Discovery project, complete the following steps:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **LDAP directory**, and then click **Next**.
4. Name the collection.
5. If the language of the documents in Salesforce is not English, select the appropriate language.

   For a list of supported languages, see Language support.
6. **Optional**: Change the synchronization schedule.

   The crawler schedule options work as follows for LDAP directories:

Full crawling

Crawls all entries.

Crawling updates

Crawls all entries, then filters out any entries that were inserted, updated, or deleted since the last crawl.

Crawling new and modified content

Runs an LDAP query against the data source server to pick up any entries that were inserted or updated only.

For more information, see Crawl schedule options.

7. Configure a secure connection to the directory.

Server type

Choose your server type from the following options:

- IBM Security Directory Server
- Microsoft Active Directory
- Oracle Directory Server

LDAP protocol

If you want to encrypt data and verify the server certificate over Transport Layer Security (TLS), choose `ldaps`.

LDAP host name

Specify the hostname of the directory server. For example: `<ldap-hostname>.mydomain.com`.

LDAP host port

By default, the LDAP port is `389` and the LDAP-S port is `636`.

LDAP binding username

If the directory server requires credentials, the username that is used to bind to the directory service.

In most cases, this username is a distinguished name (DN). The username is case-sensitive.

LDAP binding user password

The password that is associated with the username.

8. Specify the information that you want to index from the directory.

LDAP Base DN

The object where you want to start the crawl.

LDAP directories have a hierarchical tree structure of objects. The base search distinguished name specifies the subtree in which you want the crawl to be constrained.

DN is a *distinguished name* that is defined by a series of *relative distinguished names* separated by commas. Each relative distinguished name consists of an *attribute* name-and-value pair that represents an object in a directory.

For example, in Active Directory, attributes can include a common name (CN) such as `Jane Doe` and an organizational unit (OU) such as `Research`. Most distinguished names include one or more domain component (DC) attributes, which define the namespace where the LDAP directory is hosted.

Here's an example of a distinguished name for Jane:

```
CN=Jane Doe,OU=Research,DC=IBM,DC=COM
```

LDAP user filter

A filter to apply to the search to use to find LDAP entries that you want to crawl.

If unspecified, a default value is applied that is considered the best filter for the server type that you selected. You can edit the predefined filter value.

- Expand the *Advanced configuration* section to list specific attributes to include or exclude from the search.

  For example, you might need to know the country in which an employee works, so you want to include a `c` attribute that stores the ISO country code. Or maybe you never want to return an employee's serial number, so you exclude the `serialnumber` attribute.

- Specify the search scope. You can choose to crawl records that are one level from the search base DN or to crawl the entire subtree that is associated with the search base DN.

- If the LDAP directory data source has binary attributes, you can enable the **Allow binary attributes** option.

  When enabled, the crawler creates a separate document for each binary attribute that is specified. The document also contains any other non-binary LDAP attribute values.

  For more information about the binary option, see RTF 4522.

  In the **Binary attributes** field, specify the names of the binary attributes that you want to index.

9. If you want the crawler to extract text from images on the site, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

   > **Note:** When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.

10. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Local File System

Crawl documents that are stored in a local file system.

IBM Cloud Pak for Data **IBM Cloud Pak for Data only**

> **Note:** This information applies only to installed deployments.

What documents are crawled

- Only file types that are supported by Discovery in your file path are crawled; all others are ignored. For more information, see Supported file types.
- Only files in the `/mnt` directory or one of its subdirectories can be accessed by the crawler.
- Only files with file extensions that match the file extension filter rules that you specify are crawled. *Added with the 4.7.0 release.*
- When a source is recrawled, new documents are added, updated documents are modified to the current version, and deleted documents are deleted from the collection's index.
- All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

Prerequisite steps

Before you connect to the Local File System data source, complete the following step:

- Create a persistent volume claim on the crawler pod

The service uses Portworx storage by default. However, if you are using Network File System (NFS) storage, see Prerequisite steps for NFS storage instead.

Creating and mounting a persistent volume claim on the crawler pod

Before you can crawl a local file system, you must create a persistent volume claim and mount it on the `crawler` pod. You also need to copy the files that you want to crawl to the Discovery cluster that you are working on. If you have multiple Discovery clusters, you must copy the files along with the `crawler-pvc-portworx.yaml` file that you will create in this task to each cluster.

Complete the following steps:

1. Enter the following command to check the `storageclass` name of the Portworx provisioner:

   ```
   $ oc get storageclass | grep portworx-gp3-sc
   ```

   You might see output similar to the following:

   ```
   NAME             PROVISIONER                    RECLAIMPOLICY  VOLUMEBINDINGMODE  ALLOWVOLUMEEXPANSION  AGE
   portworx-gp3-sc  kubernetes.io/portworx-volume  Retain         Immediate          true                  51d
   ```

2. Create a file named `crawler-pvc-portworx.yaml` to define the persistent volume claim (PVC) with the following content:

   ```
   kind: PersistentVolumeClaim
   apiVersion: v1
   metadata:
     name: <name-of-portworx-pvc>
   spec:
     accessModes:
       - ReadWriteMany
     resources:
       requests:
         storage: 10Gi
     storageClassName: portworx-gp3-sc
   ```

   Replace `<name-of-portworx-pvc>` with the name of your dynamic Portworx persistent volume claim. For example, `jdoe-pvc-portworx`

3. Enter the following command to create the persistent volume claim:

   ```
   $ oc create -f crawler-pvc-portworx.yaml
   ```

   A message is displayed:

   ```
   persistentvolumeclaim/jdoe-pvc-portworx created
   ```

4. Enter the following command to mount the persistent volume claim to the `crawler` pod:

   ```
   $ oc patch wd wd --type=merge \
   --patch='{"spec": {"ingestion": {"crawler": {"mount": {"enabled": true, "persistentVolumeClaimName": "<name-of-portworx-pvc>" } } } } }'
   ```

   Replace `<name-of-portworx-pvc>` with the name of your dynamic Portworx persistent volume claim. For example, `jdoe-pvc-portworx`.

5. Enter the following command to copy the files that you want to crawl to your dynamic Portworx persistent volume claim.

   You only need to run this command one time against one of the existing `crawler` pods. The persistent volume claim is shared among all `crawler` and `ingestion-api` pods. Replace the variables in the command with the appropriate information.

   ```
   $ oc rsync <path-to-local-file-system-folder> <crawler-pod>:/mnt
   ```

You mounted the persistent volume claim (PVC) and copied the files that you want to crawl to the PVC.

Connecting to a local file system data source

From your Discovery project, complete the following steps:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **Local File System**, and then click **Next**.
4. Name the collection.
5. If the language of the documents that you want to crawl is not English, select the appropriate language.

   For a list of supported languages, see Language support.
6. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.
7. In the *Specify what you want to crawl* section, enter the file path that you want to crawl in the **Path** field, and then click **Add**.

The file path is case-sensitive. Remember, only files in the `/mnt` directory or one of its subdirectories can be accessed by the crawler.

8. Optionally, add more file paths.
9. If you want to limit the types of files to add to the collection, you can list the file extensions for file types to either include or exclude.

   For a list of supported file types, see Supported file types.

   > **Note:** Support for this option was added with the 4.7.0 release.

10. If you want the crawler to extract text from images in documents, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

    > **Note:** When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.

11. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Prerequisite steps for NFS storage

Choose one of the following methods to enable the `crawler` pod to access the file system:

- Configure an external NFS server
- Configure dynamic provisioning with an NFS storage class

Configuring an external NFS server

If the local file system files or folders that you want to crawl are stored in an external Network File System (NFS), you can use the external NFS server to create the persistent volume claim.

1. Create a file named `crawler-pv-nfs.yaml` with the following content:

```
apiVersion: v1
kind: PersistentVolume
metadata:
  name: <persistent-volume-name>
  labels:
    pv-name: <persistent-volume-name>
spec:
  capacity:
    storage: 10Gi
  accessModes:
    - ReadWriteMany
  persistentVolumeReclaimPolicy: Retain
  nfs:
    server: <NFS server hostname or IP address>
    path: <Path of NFS exported folder>
```

   Replace references to `<persistent-volume-name>` with the name of your persistent volume. For example, `jdoe-nfs-pv` and add the missing external NFS details.

2. Enter the following command to create the persistent volume claim:

```
$ oc create -f crawler-pv-nfs.yaml
```

   The following message is displayed:

```
persistentvolume/jdoe-nfs-pv created
```

3. Create a file called `crawler-pvc-nfs.yaml` with the following content:

```
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: <persistent-volume-claim-name>
spec:
  accessModes:
```

```
    - ReadWriteMany
  resources:
    requests:
      storage: 10Gi
  selector:
    matchLabels:
      pv-name: <persistent-volume-name>
```

Replace the following variables:

- `<persistent-volume-claim-name>`: Specify the name of your persistent volume claim. For example, `jdoe-nfs-pvc`.
- `<persistent-volume-name>`: Specify the name of your persistent volume. For example, `jdoe-nfs-pv`.

4. Enter the following command to create the persistent volume claim:

```
$ oc create -f crawler-pvc-nfs.yaml
```

The following message is displayed:

```
persistentvolumeclaim/jdoe-nfs-pvc created
```

5. Enter the following command to mount the persistent volume claim to the `crawler` pod.

This command also mounts the persistent volume claim to all `ingestion-api` pods. Replace `<persistent-volume-claim-name>` with the name of your persistent volume claim. For example, `jdoe-nfs-pvc`.

```
$ oc patch wd wd --type=merge \
--patch='{"spec": {"ingestion": {"crawler": {"mount": {"enabled": true, "persistentVolumeClaimName": "<persistent-volume-claim-name>" } } } }
}'
```

Configuring dynamic provisioning with an NFS storage class

If you want to crawl your local file system files or folders but you do not want to prepare an extra NFS server to store those files or folders, you can configure dynamic storage by using an NFS storage class.

For more information about storage providers that Discovery supports and for storage comparisons, see Storage considerations.

Before you complete this task, copy the files that you want to crawl to the Discovery cluster that you are working on. If you have multiple Discovery clusters, you must copy the files along with the `crawler-pvc-dynamic.yaml` file that you create in this task to each cluster.

Complete the following steps:

1. Enter the following command to check the `storageclass` name of the NFS provisioner:

```
$ oc get storageclass
```

A message is displayed.

```
NAME         PROVISIONER                                    RECLAIMPOLICY  VOLUMEBINDINGMODE
ALLOWVOLUMEEXPANSION  AGE
nfs-client   cluster.local/innocence-nfs-client-provisioner  Delete         Immediate          true
177m
```

2. Create a file that is named `crawler-pvc-dynamic.yaml` and add the following content to it:

```
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: <name-of-dynamic-pvc>
spec:
  accessModes:
    - ReadWriteMany
  resources:
    requests:
      storage: 10Gi
  storageClassName: nfs-client
```

Replace `<name-of-dynamic-pvc>` with the name of your dynamic NFS persistent volume claim. For example, `jdoe-dynamic-pvc`.

3. Enter the following command to create the persistent volume claim:

```
$ oc create -f crawler-pvc-dynamic.yaml
```

A message is displayed.

```
persistentvolumeclaim/jdoe-dynamic-pvc created
```

4. Enter the following command to mount the persistent volume claim to the `crawler` pod.

   This command also mounts the persistent volume claim to all `ingestion-api` pods.

   ```
   $ oc patch wd wd --type=merge \
   --patch='{"spec": {"ingestion": {"crawler": {"mount": {"enabled": true, "persistentVolumeClaimName": "<name-of-dynamic-pvc>" } } } } }'
   ```

   Replace `<name-of-dynamic-pvc>` with the name of your dynamic NFS persistent volume claim in the previous step. For example, `jdoe-dynamic-pvc`.

5. Enter the following command to copy the files that you want to crawl to your dynamic NFS persistent volume claim.

   You must run this command only one time against one of the existing `crawler` pods. The persistent volume claim is shared among all `crawler` and `ingestion-api` pods. Replace the variables in the command with the appropriate information.

   ```
   $ oc rsync <path-to-local-file-system-folder> <crawler-pod>:/mnt
   ```

You mounted the persistent volume claim (PVC) and copied all of the files that you want to crawl to the PVC.

HCL Notes

Crawl an HCL Notes (formerly Lotus Notes) database.

IBM Cloud Pak for Data

Note: This information applies only to installed deployments.

What documents are crawled

- Each document in the HCL Notes database is crawled and added to the collection as a document.
- If an HCL Notes document has a file attachment, and you choose to process file attachments, only documents that are supported by Discovery are crawled; all others are ignored. For more information, see Supported file types.
- If you choose to process attachments, the crawler attempts to crawl and index files that are attached to HCL Notes documents. File types that are supported by Discovery are indexed. For more information, see Supported file types.
- Document-level security is supported. When this option is enabled, your users can crawl and query the same content that they can access when they are logged in to HCL Notes. For more information, see Supporting document-level security.
- When a source is recrawled, new documents are added, updated documents are modified to the current version, and deleted documents are deleted from the collection's index.
- All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

Data source requirements

In addition to the data source requirements for all installed deployments, your HCL Notes data source must meet the following requirements:

- The data source can crawl HCL Notes 9.0.1 databases.
- The HCL Notes data source supports the Domino Internet Inter-ORB Protocol (DIIOP) protocol only.
- To crawl documents, including ACLs, you must have at least `Reader` level access to server, database, and document access on the Domino server.
- For group extractions from the internal Domino LDAP directory, you must have `Reader` access to the `names.nsf` directory database.
- For group extractions from the external LDAP directory, you must have the credential for the external LDAP server.

Prerequisite steps

- If you want to enable document-level security, you must take some steps to set it up. For more information, see Supporting document-level security.

  You can use the LDAP server that is used by HCL Notes (either the internal Domino LDAP or an external LDAP directory) as a remote LDAP directory to manage document-level security. Users who search the collection can be listed in an external LDAP directory. However, the user credentials that you use to set up the crawl must belong to a user who is listed in the internal Domino LDAP directory.

  To configure document-level security, you need to collect the following information:

  LDAP server URL

The LDAP server URL to connect to. For example, `ldap://<ldap_server>:<port>`.

LDAP binding username

The username to use to bind to the directory service. This user must have administrative access and be listed in the internal Domino LDAP directory.

LDAP binding user password

The password that is associated with the user.

LDAP base DN

The starting point for searching user entries in LDAP. For example, `CN=Users,DC=example,DC=com`.

LDAP user filter

The filter to apply to searches for user entries in LDAP. If unspecified, the default value is `(userPrincipalName=\{0\})`.

LDAP group filter

The filter to apply to searches for group entries in LDAP.

- Before you can crawl servers by using the Domino Internet Inter-ORB Protocol (DIIOP) protocol, you must configure the HCL Notes server to use the protocol. The server that you want to crawl must be running the DIIOP and HTTP tasks.

To configure the HCL Notes server to use DIIOP, complete the following steps:

1. Configure the HCL Notes server document.
   - In HCL Notes, open the `server` document on the HCL Notes server that you want to crawl. This document is stored in the Domino directory.
   - On the Configuration page, expand the *server* section.
   - On the Security page in the Programmability Restrictions section, specify the appropriate security restrictions for your environment in the following three fields:
     - **Run restricted Lotus Script/Java agents**
     - **Run restricted Java/Javascript/COM**
     - **Run unrestricted Java/Javascript/COM**

     For example, you might specify an asterisk (`*`) to allow unrestricted access by LotusScript/Java agents and specify usernames that are registered in the Domino directory for the Java/JavaScript/COM restrictions.

     ⚠ **Important:** To crawl a server that uses the DIIOP protocol, your configured crawler must be able to access the usernames that you specify in these fields.
   - Open the Internet Protocol page, and then open the HTTP page. Set the **Allow HTTP clients to browse database** option to **Yes**.
2. Configure the user document.
   - Open the `user` document for the user whose credentials that you want to use for LDAP binding. This document is stored in the Domino directory.
   - On the Basics page in the **Internet password** field, specify a password.

     You specify this user and password information when you set up the data source.
3. Restart the DIIOP task on the HCL Notes server.

For more information, see Running server tasks in the HCL Notes documentation.

Connecting to an HCL Notes data source

From your Discovery project, complete the following steps:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **Notes**, and then click **Next**.
4. Name the collection.
5. If the language of the documents in HCL Notes is not English, select the appropriate language.

For a list of supported languages, see Language support.

6. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.

7. In the *Enter your credentials* section, add values to the following fields:

   Host name

   The hostname of the HCL Notes server.

   User name

   The username to use to crawl the HCL Notes server.

   Password

   The password that is associated with the user.

8. In the *Crawl type*, choose what you want to crawl from the following options:
   - If you want to crawl a specific HCL Notes database, choose **Database**, and then add the file name of the database to the **Database file name** field.
   - If you want to crawl multiple databases, choose **Directory**. Specify the directory in which the databases that you want to crawl are stored in the **Directory name** field.

9. **Optional**: In the *Security* section, specify whether you want to enable document-level security.
   - If you want to enable document-level security, set the **Enable Document Level Security** switch to `On`.

     When set to **On**, your users can crawl the same content that they have access to in a HCL Notes database or directory.

   - To use the Domino LDAP directory, set the **Use remote LDAP directory** switch to `On`. Provide details about the Domino LDAP directory. You collected this information when you performed the prerequisite step.

     LDAP server URL

     The LDAP server URL to connect to. For example, `ldap://<ldap_server>:<port>`.

     LDAP binding username

     The username to use to bind to the directory service.

     LDAP binding user password

     The password that is associated with the user.

     LDAP base DN

     The starting point for searching user entries in LDAP. For example, `CN=Users,DC=example,DC=com`.

     LDAP user filter

     The filter to apply to searches for user entries in LDAP. If unspecified, the default value is `(userPrincipalName=\{0\})`.

     LDAP group filter

     The filter to apply to searches for group entries in LDAP.

10. **Optional**: In the *Advanced options* section, make choices about the following configuration settings:

    Crawl attachments

    If you want to crawl files that are attached to HCL Notes documents, set the switcher to `On`.

    Automatic code page detection

    If you want the encoding converter to detect the code of pages to crawl, keep the switch set to `On`. If you set the switcher to `Off`, specify values for the following fields:

Code page to use

Specify the character encoding of the pages that you want to crawl. If unspecified, the default value of `UTF-8` is used.

Notes formula

Specify a HCL Notes formula to use to filter the data that you want to crawl. For example, `SELECT @IsAvailable(Year) & Year > 2003`.

For more information, see Formula language in the HCL Notes documentation.

11. Specify the date that you want to use when you filter the documents. The date is stored in a field that is named `_ _$Date$_ _` in HCL Notes documents. By default, the field stores the last modified date of the document. You can choose a different date to store in the field instead.

    Document modification date

    Uses the date that the document was last modified. This option is selected by default.

    Document crawl date

    Uses the last crawled date.

    Document creation date

    Uses the creation date of the document.

12. If you want the crawler to extract text from images in documents, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

    > **Note:** When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.

13. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Salesforce

Crawl documents that are stored in a Salesforce data source.

IBM Cloud **IBM Cloud only**

> **Note:** This information applies only to managed deployments. For more information about connecting to Salesforce from an installed deployment, see Salesforce.

What documents are crawled

During the initial crawl of the content, documents from all of the objects that can be accessed from the URL that you specify are crawled and added to your collection. Knowledge Articles are crawled only if their **version** is `published` and their languages is `en-us`.

During subsequent scheduled recrawls, only new and modified documents are crawled and any changes are reflected in your collection. Documents that are deleted from the external data source are not deleted from the collection.

All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

Discovery can crawl the following objects:

- Any default and custom objects that you have access to
- Accounts
- Contacts
- Cases

- Contracts
- Knowledge articles
- Attachments

Data source requirements

In addition to the data source requirements for all managed deployments, your Salesforce data source must meet the following requirements:

- The instance that you plan to connect to must be part of an Enterprise plan or higher.
- You must obtain any required service licenses for the data source that you want to connect to. For more information about licenses, contact the system administrator of the data source.

What you need before you begin

You must have the following information ready. If you don't know it, ask your Salesforce administrator to provide the information or consult the Salesforce developer documentation.

Username

The `username` of an account that has access to the Salesforce site. For example, `jdoe@example.com`

Password

The password associated with the username. For example, `myP@ssw0rd`.

Service token

A valid Salesforce security token. For example, `mnaO8jsRET5CiJww9JnURlNN`.

URL

The URL of the Salesforce site that you want to crawl. For example, `https://my.salesforce.com`

Connecting to the data source

To configure the Salesforce data source, complete the following steps in Discovery:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **Salesforce**, and then click **Next**.
4. Add values to the following fields:
   - Username
   - Password plus service token

     To form the password, concatenate the Password and Service token values that you noted earlier. For example, `myP@ssw0rdmnaO8jsRET5CiJww9JnURlNN`. The password and token values are never returned and are used only when credentials are created or modified.
   - URL

   Click **Next**.
5. Name the collection.
6. If the language of the documents on the site is not English, select the appropriate language.

   For a list of supported languages, see Language support.
7. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.
8. Select the objects that you want to crawl.

   The more objects that you select, the longer the processing of the documents takes.
9. If you want to limit the types of files to add to the collection, you can list the file extensions for file types to either include or exclude.

   ⚠ **Important:** When you choose to list extensions for file types to exclude, you must add at least one file extension.

For a list of supported file types, see Supported file types.

10. If you want the crawler to extract text from images on the site, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

    Note: When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.

11. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Microsoft SharePoint Online

Crawl documents that are stored in a Microsoft SharePoint Online data source.

IBM Cloud **IBM Cloud only**

Note: This information applies only to managed deployments. For more information about connecting to SharePoint Online from an installed deployment, see SharePoint Online.

What documents are crawled

During the initial crawl of the content, documents from all of the objects that can be accessed from the site collection path that you specify are crawled and added to your collection. You cannot limit the crawl to one library within a site collection, for example. All objects in the specified Site collection path are crawled. Custom metadata that is associated with the SharePoint content is crawled also. You can crawl one site collection path per collection. You cannot crawl *Personal SiteCollections*.

During subsequent scheduled recrawls, only new and modified documents are crawled and any changes are reflected in your collection. Documents that are deleted from the external data source are not deleted from the collection.

All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

Discovery can crawl the following objects:

- SiteCollections
- Sites
- SubSites
- Lists
- List Items
- Document Libraries
- List Item Attachments

Data source requirements

In addition to the data source requirements for all managed deployments, your SharePoint Online data source must meet the following requirements:

- The Site Collection that you connect to must be one that was created with an Enterprise plan. It cannot be a collection that was created with a frontline worker plan.
- You must have an Azure Active Directory user ID with permission to read all of the objects that you want to crawl. For example, `<admin_user>@.onmicrosoft.com`. The user ID does not need `SiteCollection Administrator` permission.

You can choose how to authenticate with the external Microsoft SharePoint account from the following options:

Open Authentication (OAuth v2)

Authenticates with the external data source by using a token so that your user credentials do not need to be shared. With this authentication method, you can log in to your Microsoft account directly to generate a token that is used by Discovery to connect to your data.

The *Sign in with Microsoft* option that uses Open Authentication v2 to authenticate with the external data source is a beta feature.

Before anyone can create connectors that use this authentication method, a user with the *Global Administrator* role must complete a one-time prerequisite steps to authorize the connection for all projects in the Discovery service instance.

Security Assertion Markup Language (SAML)

An older mechanism for authentication and authorization that requires user credentials to be shared with the Discovery service.

If you choose to use this authentication method, your Microsoft SharePoint account must meet the following requirements:

- Unless you created your SharePoint Online account before January 2020, two-factor authentication is enabled for the account by default. You must disable two-factor authentication.

  To view and change your multifactor authentication status, see View the status for a user or Change the status for a user.

- The crawl user account must have legacy authentication and `Contribute` level permissions enabled.

  To enable legacy authentication, go to the Azure portal or contact your SharePoint administrator.

- The connector supports the `Password hash synchronization (PHS)` method for enabling hybrid identity only. Use any other type (such as Pass-through authentication or Federation) at your own risk.

- You must know the following information:

  Username

  The username of the user account to use to connect to the SharePoint Online SiteCollection that you want to crawl.

  For example, `<janedoe>@exampledomain.onmicrosoft.com`.

  Password

  The password to connect to the SharePoint Online SiteCollection that you want to crawl.

  This value is never returned and is only used when credentials are created or modified.

What you need before you begin

You must have the following information ready. If you don't know it, ask your SharePoint administrator to provide the information or consult the Microsoft SharePoint developer documentation:

Organization URL

The root URL of the source that you want to crawl. Specify the domain name of the URL, for example `https://<company>.<domain>.com`.

Site collection path

The `site_collection_path` to the section of the site where you want to start the crawl.

For example, if the content that you want to crawl is available from `https://<company>.<domain>.com/sites/test`, then you can specify `https://<company>.<domain>.com` as the Organization URL and `/sites/test` as the Site collection path.

- You cannot specify folder paths as input.
- You cannot specify a path to an Active Server Page Extended (ASPX) file, such as URLs to document libraries, lists, and subsites.
- If you don't specify a path, the default value of `/` is used, and the root site collection is crawled.

- **Application ID**: ID of the data source that you want to crawl. This information is required only if you want to store ACL information that is associated with the source documents.

One-time prerequisite step for OAuth

Before anyone can configure the connector to use OAuth v2 authentication method, a user with the *Global Administrator* role in Microsoft Azure Directory where the data source is located must complete steps to register the Discovery enterprise application in Microsoft Azure. This step must be completed once per Discovery service instance.

The administrator does not need to create the application in Azure. When they choose SharePoint Online as the data source, the Discovery service generates the app automatically. As described in the procedure to follow, during the set up of the connector, the administrator must log in to Microsoft with credentials for a user with the *Global Administrator* role in Microsoft Azure Directory and allow the enterprise application to be registered.

The following steps must be completed by a global administrator one time only per service instance:

1. Review the default user access settings that will be applied to the enterprise application in Microsoft Azure.

Enterprise applications can handle user access in many ways. Check the default settings to ensure that they are appropriate for your deployment by completing the following steps:

1. Log in to Microsoft Azure.
2. From the *Enterprise applications* page in *Azure Active Directory*, click *Consent and permissions*.

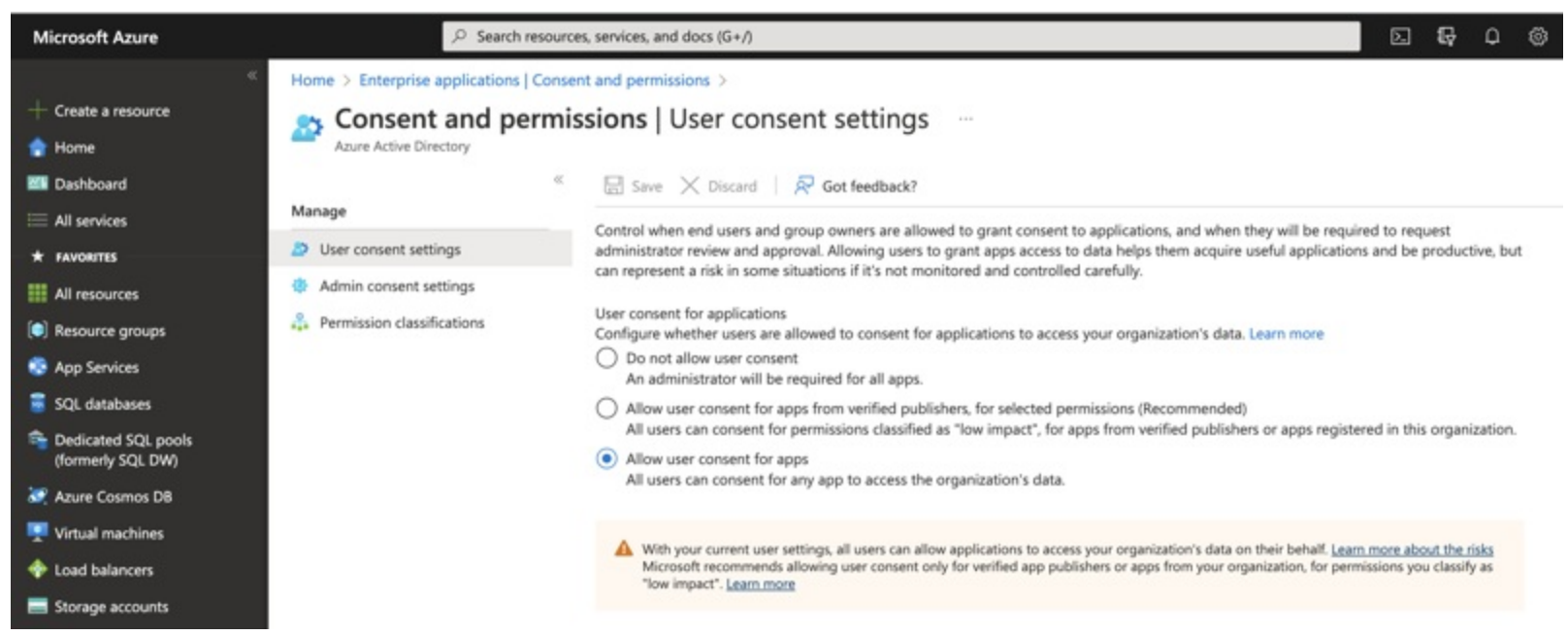

Figure 1. Microsoft Azure Enterprise application permissions user interface

1. Do one of the following things:
   - If *Allow user consent for apps* is selected, no more action is needed.
   - If *Allow user consent for apps from verified publishers, for selected permissions* is selected, then complete the following steps:

     Click *Permissions classifications* link, and then ensure that the following permissions are configured at a minimum:

     - Office 365 SharePoint Online: MyFiles.Read
     - Office 365 SharePoint Online: AllSites.Read
     - Microsoft Graph: offline_access
     - Microsoft Graph: profile

   The *Do not allow user consent option* is not supported.

   The settings that you specify will be applied to the enterprise application that is created by Discovery in subsequent steps.
2. From the navigation pane of Discovery, choose **Manage collections**.
3. Click **New collection**.
4. Click **SharePoint Online**, and then click **Next**.
5. Add a URL to the **Organization URL** field.
6. Click **Sign in with Microsoft**.

   Pop-ups must be enabled for this site in your web browser.

   The *Sign in with Microsoft* option that uses Open Authentication to authenticate with the external data source is a beta feature.

   Log in to your Microsoft SharePoint account with your user name and password, and then complete two-factor authentication, if necessary.

   > ⚠ **Important:** Remember, the credentials you use must have the *Global Administrator* role in Microsoft Azure Directory. If you are not prompted for a user name and password, take note. You might be logged in to a Microsoft Sharepoint account already. If you are logged in to an account that you don't want to use for this connector, stop here. (Any account where you are logged in will be used automatically. And you cannot change the account configuration later.) Open a web browser in incognito mode and start this procedure over from step 1.

   Discovery generates an enterprise application that it will register with the SharePoint organization that you specify. The enterprise application name has the format *IBM App Connect_{unique name}*.
7. Review the permissions that are associated with the enterprise application that Discovery will register, and then select **Consent on behalf of your organization**.

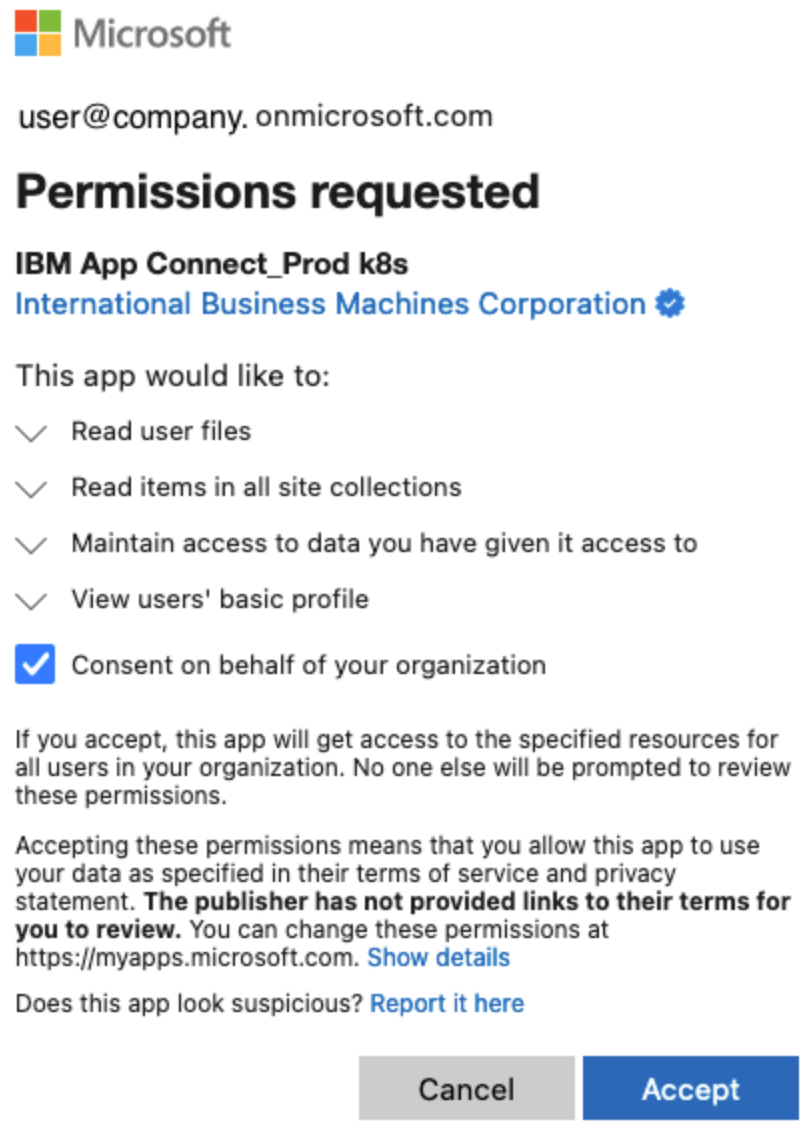

Figure 2. Discovery permission request dialog

8. Click **Accept**.
9. If you want to create a collection, you can name the collection, and then click **Finish**.

   Otherwise, you can click **Back** to exit the collection creation process.

Now, anyone from your organization who works in a project that is hosted by the same Discovery service instance can create a collection by using the SharePoint Online connector.

OAuth support revisions

Support for the OAuth method of authentication was added with a software update in February 2022. If you want to update an existing connector to use OAuth instead of SAML, you must re-create the connector. You cannot change the authentication mechanism for an existing connector.

The OAuth method of authentication was updated in January 2023. The enterprise application that is registered with Microsoft Azure now requires *Read* access only. Previously, the enterprise application required *Write* access. If you want to take advantage of this change, delete your current enterprise application and recreate the connector. For more information about how to delete an enterprise application, see the Microsoft documentation.

Connecting to the data source

To configure the Microsoft SharePoint Online data source, complete the following steps in Discovery:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **SharePoint Online**, and then click **Next**.
4. Add a URL to the **Organization URL** field.
5. To enable access to your external data source, choose the method that you want to use to authenticate with the data source from the following options:

   Open Authentication (OAuth v2)

   Click **Sign in with Microsoft**.

   Pop-ups must be enabled for this site in your web browser.

   The *Sign in with Microsoft* option that uses Open Authentication to authenticate with the external data source is a beta feature.

   Log in to your Microsoft SharePoint account with your user name and password, and then complete two-factor authentication, if necessary.

Security Assertion Markup Language (SAML)

Specify a username and password for a user that is authorized to access the site you want to crawl, and then click **Next**.

6. Specify the path you want to crawl in the **Site collection path** field.
7. Name the collection.
8. If the language of the documents on the site is not English, select the appropriate language.

   For a list of supported languages, see Language support.
9. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.
10. **Optional**: If you want to store any access control information that exists in the SharePoint documents that you crawl, in the *Security* section, set the **Include Access Control List** switch to `On`.

    When you enable this option, information about SharePoint access rules that is stored in SharePoint source documents is retained and stored as metadata in the documents that are added to your collection.

    This feature is not the same as enabling document-level security for the collection. The access rules in the document metadata are not used by Discovery search. Enabling this feature merely stores the information so that you can leverage the access rules when you build a custom search solution.

    ⚠ **Important:** Use of this feature increases the size of the documents that are generated in the collection and increases the crawl time. Only enable the feature if your use case requires that you store the SharePoint document ACL information.

    If you enable this feature, someone with the administrator role in Microsoft SharePoint must take extra steps to ensure that users who crawl the site have the right permissions to access ACL metadata.

    An administrator must complete the following steps:

    1. Log in to Microsoft SharePoint.
    2. Open the page for your SharePoint site.
    3. From the settings menu, choose *Site permissions*.
    4. Click *Advanced permission settings*.
    5. Make sure that people who want to collect access control information during a crawl have or are members of a group that has the *Full Control* permission for the site.

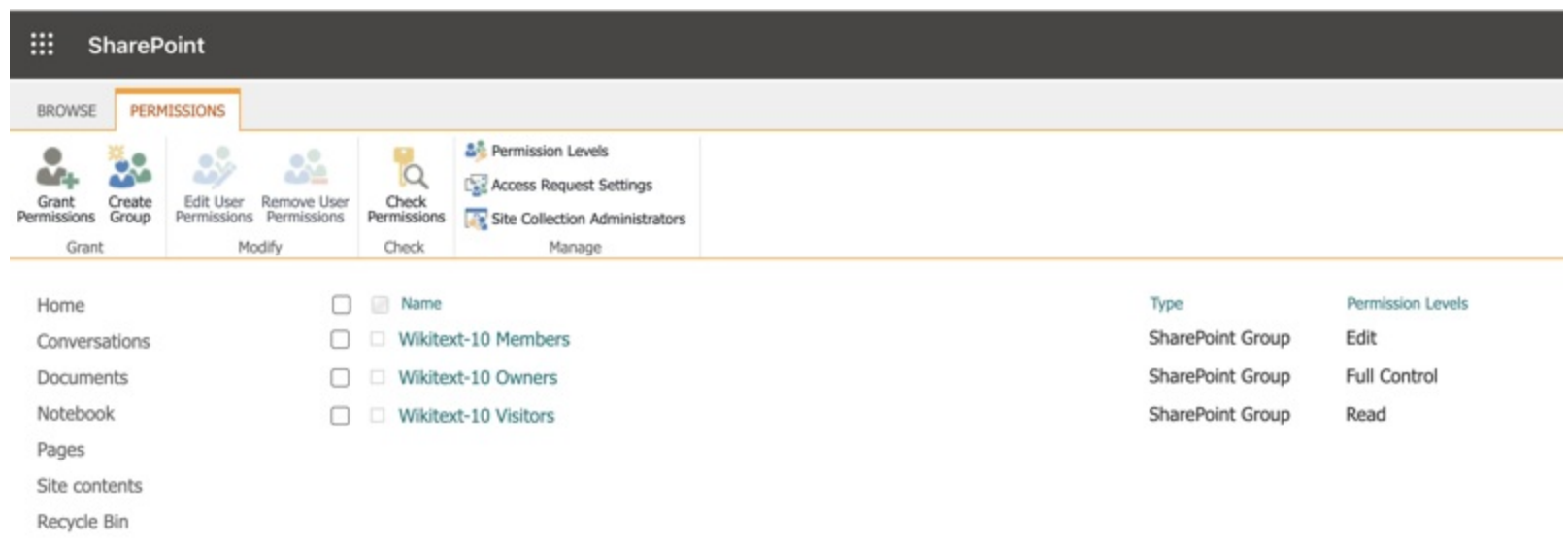

Figure 3. Microsoft SharePoint permissions user interface

Note: When access control list information is not extracted, *Read* permission is sufficient for all users who crawl the content.

11. If you want to limit the types of files to add to the collection, you can list the file extensions for file types to either include or exclude.

    ⚠ **Important:** When you choose to list extensions for file types to exclude, you must add at least one file extension. By default, the *Extension filter* is applicable to SharePoint *Document Libraries* and *List Item Attachments* objects only. To apply the filter to all SharePoint object types, set **Apply extension filter to all SharePoint object types** to **On** on the user interface.

    For a list of supported file types, see Supported file types.

12. If you want the crawler to extract text from images on the site, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

    Note: When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.

13. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Note: You cannot currently change the user account that is associated with the OAuth setup later, nor any of the details of the existing user account that the connector is configured to use. For example, you cannot update the password that was used to set up the connection after a password change in SharePoint.

Sample access control list information

The following screen capture illustrates the type of ACL information that is stored in the document when you include the access control list.

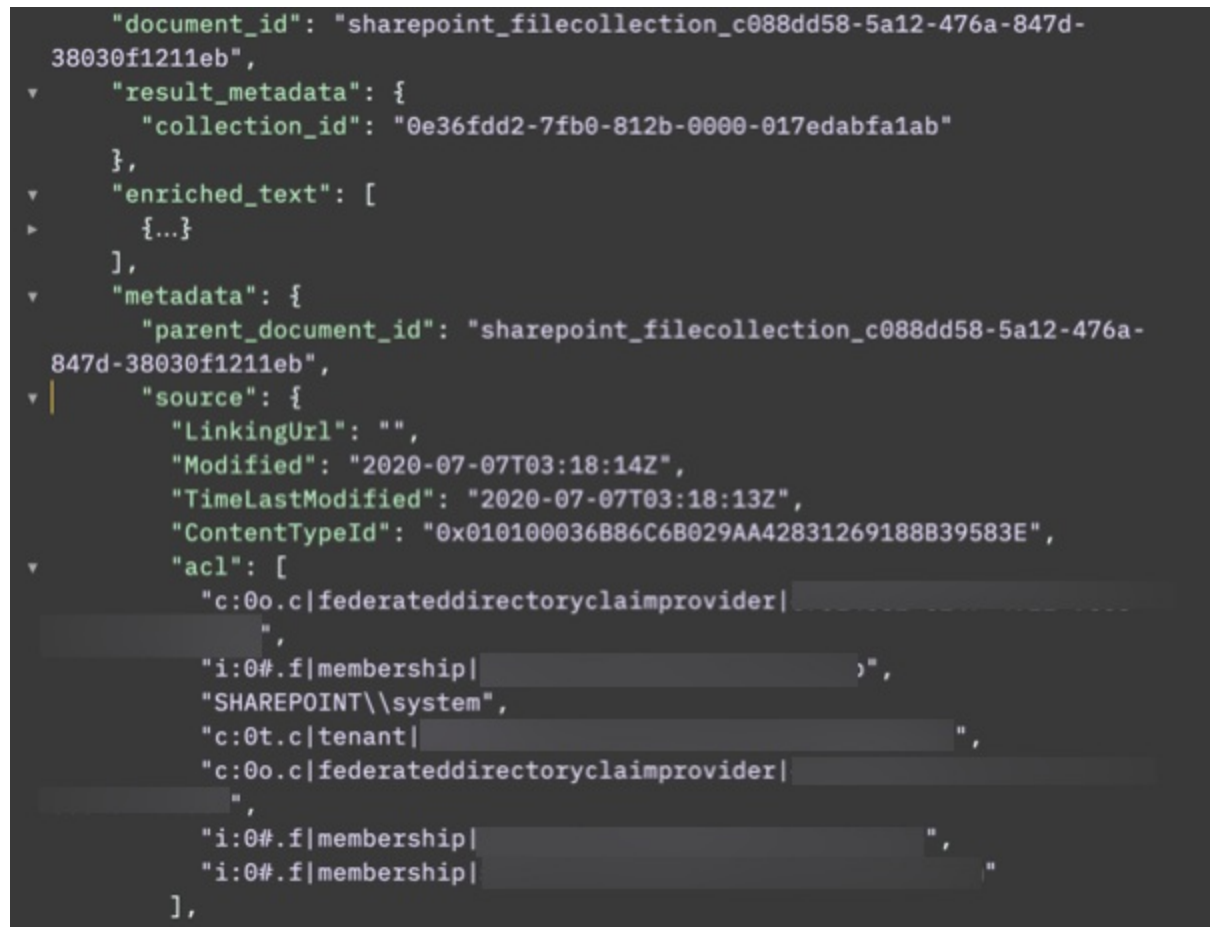

Figure 4. Representation of ACL information in document metadata

Microsoft SharePoint On Prem

Crawl documents that are stored in a Microsoft SharePoint data source that is hosted on premises.

IBM Cloud **IBM Cloud only**

Note: This information applies only to managed deployments. For more information about connecting to an on-premises SharePoint data source from an installed deployment, see SharePoint On Prem.

What documents are crawled

During the initial crawl of the content, documents from all of the objects that can be accessed from the site collection path that you specify are crawled and added to your collection. Custom metadata that is associated with the SharePoint content is crawled also. You can crawl one site collection path per collection. You cannot crawl *Personal SiteCollections*.

During subsequent scheduled recrawls, only new and modified documents are crawled and any changes are reflected in your collection. Documents that are deleted from the external data source are not deleted from the collection.

All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

The following table illustrates the objects that Discovery can crawl.

Data source	Objects that are crawled

Microsoft SharePoint On Prem	SiteCollections, Sites, SubSites, Lists, List Items, Document Libraries, List Item Attachments

Table 1. Data sources crawling support

Data source requirements

In addition to the data source requirements for all managed deployments, your SharePoint On Prem data source must meet the following requirements:

- You can connect to a SharePoint 2013, 2016, or 2019 on-premises data source.
- The user ID must have `SiteCollection Administrator` permission and be able to access all of the sites and lists that they want to crawl.
- The crawler supports Windows New Technology LAN Manager (NTLM) v1 authentication only. It does not support NTLM v2 or Security Assertion Markup Language (SAML) authentication.

What you need before you begin

You must have the following information ready. If you don't know it, ask your SharePoint administrator to provide the information or consult the Microsoft SharePoint developer documentation:

Username

The username to use to connect to the SharePoint On Prem web application that you want to crawl. For example, `siteadmin01`.

Password

The password to connect to the SharePoint On Prem web application that you want to crawl. This value is never returned and is only used when credentials are created or modified.

Web Application URL

The SharePoint web application URL. For example, `https://sharepointwebapp.com:8443`. If you do not enter a port number, the default value of `80` is used for an HTTP URL and `443` for HTTPS.

Domain

The domain name of the SharePoint On Prem account. For example, `sharepoint.mycointernal`.

Prerequisite step

Before you can connect to a SharePoint On Prem data source, you must install and configure IBM® Secure Gateway for IBM Cloud®.

For more information, see Installing IBM Secure Gateway for on-premises data.

Connecting to the data source

To configure the Microsoft SharePoint On Prem data source, complete the following steps in Discovery:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **SharePoint On Prem**, and then click **Next**.
4. Add values to the following fields:
   - Username
   - Password
   - Web Application URL
   - Domain

   Click **Next**.
5. Name the collection.
6. If the language of the documents on the site is not English, select the appropriate language.

   For a list of supported languages, see Language support.
7. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.
8. If you want to limit the types of files to add to the collection, you can list the file extensions for file types to either include or exclude.

> **Important:** When you choose to list extensions for file types to exclude, you must add at least one file extension.

For a list of supported file types, see Supported file types.

9. If you want the crawler to extract text from images on the site, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

> **Note:** When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.

10. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Web crawl

Add a web crawl collection to crawl a website, analyze its page content, and store meaningful information. Specify one or more base web page URLs and configure how many linked pages for the web crawl to follow. You can configure how often to synchronize with the website, so you control how up to date the data in your collection is.

IBM Cloud **IBM Cloud only**

> **Note:** This information applies only to managed deployments. For more information about connecting to a website from an installed deployment, see Web crawl.

What documents are crawled

You can connect to the following types of web content:

- Public websites
- Private company websites or other sites that require authentication
- Websites that are behind a corporate firewall

During the initial crawl of the content, all website pages that match your search settings are crawled and added to the document index of your collection. The crawl starts on the web page that you specify in the *Starting URLs* field. If your collection is configured to follow links, the crawl follows links on the starting page that share the same subtree as the starting page. For example, if you specify `https://www.example.com/banking/faqs.html`, links with URLs that begin with `https://www.example.com/banking/` are crawled. If you specify `https://www.example.com/banking`, links with URLs that begin with `https://www.example.com/` are crawled.

The crawl cannot access secure subdirectories. For example, if a subdirectory that you expect the crawl to access, such as `https://www.example.com/banking/pdfs`, isn't being crawled, check whether you can access the subdirectory URL from a web browser directly. If you can't access it, the crawl can't access it.

During subsequent scheduled recrawls, a full recrawl is performed and any changes are reflected in your collection. Documents that were added to your collection from website pages that are later deleted from the external website are not deleted from the collection. However, starting with collections that were created after April 2022, when you remove a starting URL from the web crawl configuration, any associated documents are deleted. Deleted documents include indexed documents that were added to the collection based on the content of the web page at the starting URL and documents that were derived from web pages that the starting URL linked to. You cannot limit the number of indexed documents by changing other settings, such as changing the existing URL to include a path with a more limited scope than before or reducing the maximum number of links to follow to 0. Only by deleting the URL can you remove the indexed documents that are associated with it.

The web crawler can crawl web pages that use JavaScript to render content, but the crawler works best on individual pages, not entire websites. It cannot crawl sites that use dynamic URLs; if you can't see any content when you view the source code of a web page in your browser, then the service cannot crawl it.

If you want to crawl a group of URLs that includes some websites that require authentication and some that don't, consider creating a different collection for each authentication type. The connector does not support cookie-based crawling.

All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

The following table illustrates the objects that Discovery can crawl.

Objects that are crawled
Websites, website subdirectories

Table 1. Data sources crawling support

Prerequisite step

If you want to connect to a website that is hosted behind a firewall, set up an IBM® Secure Gateway for IBM Cloud® connection first.

Valuable content is often stored on your company's internal website. Typically, such intranet websites are accessible only from a computer that is connected to your office network or through a VPN connection. You can establish a persistent and more secure connection between the web crawler and this type of internal site by using Secure Gateway.

For more information about how to set up the connection, see Installing IBM Secure Gateway for on-premises data.

Connecting to the data source

To configure the web crawl collection, complete the following steps:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **Web crawl**, and then click **Next**.
4. Name the collection.
5. If the language of the content on the website is not English, select the appropriate language.

   For a list of supported languages, see Language support.
6. **Optional**: You can change the synchronization schedule.

   For more information, see Crawl schedule options.
7. Specify the URL of the website that you want to crawl.
   - If the site you want to crawl requires a login, set **Basic authentication** to `On`, add the URL of the page to the **Starting URL** field, and then click **Add**.

     Add a username and password with access to the site, and then click **Save credentials**. You can specify only one set of credentials per collection.

     For example, you can specify `https://cloud.ibm.com` as the starting URL and add your IBMid as the credentials.

     If you want to start the crawl from a specific section of the site, specify it in the **Starting URLs** field. The domain name of the subsection must match the domain in the URL you specified earlier.

     For example, you might change the starting URL to `https://cloud.ibm.com/unifiedsupport/supportcenter`.
   - For any public web pages that you want to crawl, add the URL for the root page of the website to the **Starting URLs** field, and then click **Add**. You can add more than one starting page.

     The final forward slash (`/`) in the URL determines the subtree to crawl. If you specify `https://www.example.com/banking/faqs.html`, all URLs that begin with `https://www.example.com/banking/` are crawled, for example. If you specify `https://www.example.com/banking` all URLs that begin with `https://www.example.com/` are crawled.

     By default, the number of consecutive links that the crawl follows from the starting URL is `2`. To change the number of hops or to list website sections to exclude from the crawl, click the edit icon.

     - The maximum number of hops allowed is `20`.
     - To specify URL paths to exclude, add the site path. For example, if the starting URL is `https://example.com`, you can exclude `https://example.com/pricing` by entering `/pricing/`.

       Any section of the web address that contains the site path you specify is excluded. For example, if you specify `/licenses/`, the page `https://example.com/products/licenses/europe` is excluded, among others.
     - If you want to restrict the crawl to a single page, add the URL to the **Starting URLs** field. For example, `https://www.example.com/banking/faqs.html`. Click the edit icon to set the **Maximum number of links to follow** to `0`.
   - If the website that you want to crawl uses JavaScript to customize the page content before it is displayed, you must take an extra step.

     After you enter the starting URL and click **Add**, edit the URL by clicking the edit icon. Set the *Execute JavaScript during crawl* switcher to **On**, and then click **Save**.

     Note: When JavaScript processing is enabled, it takes 3 to 4 times longer to crawl a page. Use it only on individual web pages where you know it is necessary because the page renders its content dynamically. If you see timeout messages or the crawl ends without adding content to the collection, decrease the number of web pages that are included in the crawl. For example, you can specify the

exact page to crawl in the *Starting URLs* field, and set *Maximum number of links to follow* to 0.

- To connect to a website that is hosted behind a firewall, set up an IBM® Secure Gateway for IBM Cloud® connection first.

  Expand *More connection settings*, and then set **Connect to on-premises network** to `On`. Provide details about your Secure Gateway connection.

8. Optional: Add another web address to the **Starting URLs** field.

   ⚠ **Important:** The number of starting URLs for a single collection must be less than 100. If you have a requirement to crawl a large number of websites, see I need to crawl lots of sites. What's my limit?.

   The number of web pages that are crawled is limited to 250,000, so the web crawler might not crawl all the specified websites.

   The number of child URLs per URL that are crawled is limited to 10,000. If the number of child URLs within any crawled URL exceeds 10,000, the crawler cannot process any of the content in the child URLs.

9. If you want to limit the types of files to add to the collection, you can list the file extensions for file types to either include or exclude.

   ⚠ **Important:** If the URLs for your website pages do not end in *.html*, use the exclude filter instead of the include filter. You must add at least one file extension to exclude.

   For a list of supported file types, see Supported file types.

10. If you want the web crawl to extract text from images on the site, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

    Note: When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.

11. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

I need to crawl lots of sites. What's my limit?

The service can support a total of 500 crawler connections per Discovery service instance. All of the data sources except Web crawl use one crawler connection each. For Web crawl, one connection is required for every 5 starting URLs. If you add 10 starting URLS, for example, Discovery generates the extra crawler connection that is needed to support the extra 5 URLs. Therefore, the maximum number of starting URLs that you can use depends on the other data collections that are configured in your service instance. You can calculate the limit yourself.

To calculate the starting URL limit, complete the following steps:

1. Calculate the number of other data source collections in the service instance, meaning this project and any other projects in the same Discovery instance.

   For example, you might have 2 IBM Cloud Object Store collections in one project and 2 Salesforce collections and 1 SharePoint Online collection in another project. In this example, the total number of other data source collections is 5.

2. Subtract the number of other data source collections from the maximum allowed number of crawler connections, which is 500.

   For example, 500 - 5 = 495.

3. Multiply the remainder by 5 to determine the total number of starting URLs that you can use.

   For example, 495 x 5 = 2,475.

Note: To use the maximum-allowed number of starting URLs in the example, you would need 25 web crawl collections because each collection allows a maximum of 100 starting URLs to be configured. However, don't configure your instance to use the absolute maximum number allowed. If one or more additional data sources are added subsequently to a project in this service instance, it will impact the number of starting URLs that the instance can crawl successfully.

Troubleshooting crawler issues

A 403 Forbidden error is returned

The website that you want to crawl might block requests from all but a specific set of named entities. If possible, add the crawler to the allowlist for the site. The identifying header for the crawler is `User-Agent: IBM-AppConnect/V1`.

Windows File System

Crawl documents that are stored in a Microsoft Windows file system.

IBM Cloud Pak for Data **IBM Cloud Pak for Data only**

> **Note:** This information applies only to installed deployments.

What documents are crawled

- Only documents that are supported by Discovery in your file path are crawled; all others are ignored. For more information, see Supported file types.
- Document-level security is supported. When this option is enabled, your users can crawl and query the same content that they can access when they access the file system directly.
- When a source is recrawled, new documents are added, updated documents are modified to the current version, and deleted documents are deleted from the collection's index.
- All Discovery data source connectors are read-only. Regardless of the permissions that are granted to the crawl account, Discovery never writes, updates, or deletes any content in the original data source.

Data source requirements

In addition to the data source requirements for all installed deployments, your Windows File System data source must meet the following requirements:

- The connector supports Microsoft Windows Server 2012 R2, 2016, 2019, and 2022.
- The remote agent server and the file servers to be crawled must belong to the same Windows domain. The crawler can gather access control list (ACL) data from a single Windows domain only.

> **Note:** Support for Microsoft Windows Server 2022 was added with the 4.6 release. Starting with the 4.7 release, you can secure traffic that is sent between the Windows Agent service and its crawler by enabling support for the transport layer security (TLS) protocol.

Prerequisite steps

- If you want to enable document-level security, you must take some steps to set it up. For more information, see Supporting document-level security.

  To configure document-level security, you need to collect the following information:

  LDAP server URL

  The LDAP server URL to connect to. For example, `ldap://<ldap_server>:<port>`.

  LDAP binding username

  The username to use to bind to the directory service.

  In most cases, this username is a distinguished name (DN). An Active Directory username might work, but, unlike the general Windows logon, it is case sensitive.

  LDAP binding user password

  The password that is associated with the binding username.

  LDAP base DN

  The starting point for searching user entries in LDAP. For example, `CN=Users,DC=example,DC=com`.

  LDAP user filter

  The user filter to search user entries in LDAP. If empty, the default value is `(userPrincipalName={0})`.

- Before you configure a Windows File System collection, you must install the IBM Watson Discovery Agent for Windows File Systems on a remote Windows file server or on a remote Windows server. The agent is a Windows service that retrieves data from data source servers and sends it to Discovery. The agent can crawl remote Windows file systems, drives that are local to the agent, and shared network folders.

If you install the agent on a remote Windows server, the remote Windows server must be able to mount one or more file servers so that the agent can crawl the remote Windows file systems.

To install and configure the agent, complete the following tasks:

- Install the agent.
- Configure shared directories on the agent server .
- Start and monitor the status of the agent server .

Install the agent

With the 4.6 release, the IBM Watson Discovery Agent for Windows File Systems was updated to run with 64-bit versions of Windows. If you installed the agent with a release prior to 4.6, you must uninstall the previous version, delete it, and then reinstall the agent.

Do one of the following tasks:

- You have a previous installation that is earlier than 4.6: Replace the pre-4.6 agent
- You are using the connector for the first time: Install the agent

Replace the pre-4.6 agent

Required for deployments where a version of the IBM Watson Discovery Agent for Windows File Systems that is earlier than 4.6.0.0 is installed.

To replace an earlier version of the agent, complete the following steps:

1. Copy the configuration file that defines the shared network directories that the Windows File System agent can access to a directory that is outside the agent's file path, which is `C:\Program Files (x86)\IBM\es`.

   For example, copy the `C:\Program Files (x86)\IBM\es\distributed\esadmin\config\esfsexport.txt` file to a directory such as `C:\temp` directory.

2. From the Microsoft Windows *Apps & features* utility, find the earlier version of *IBM Watson Discovery Agent for Windows File Systems* , and then click *Uninstall*.

3. Choose *Completely delete IBM Watson Discovery Agent for Windows File Systems* , and then click *Uninstall*.

4. Restart your system.

5. Complete the steps in Installing the agent to install the latest version of the agent.

6. Replace the new version of the `C:\Program Files\IBM\es\distributed\esadmin\config\esfsexport.txt` file with the file that you copied in Step 1.

   This step adds the configuration of the shared directories that you set up for the previous version of the agent to the new installation. When you reuse the file share, you can skip the step of configuring the shared directories.

7. Run the following command to verify that the directory is shared with the agent service:

   ```
   C:\Users\Administrator> esagent --lsshare
   ```

Installing the agent

To install the IBM Watson Discovery Agent for Windows File Systems for the first time, complete the following steps:

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **Windows File System**, and then click **Next**.
4. Scroll to the *Download & install Windows Agent* section, and then click **Download Windows Agent Installer**.

   A ZIP file is downloaded.

5. Decompress the `WindowsAgentServer.zip` file.
6. You can choose one of the following methods to run the installation program:
   - Double-click the `install.exe` file to launch the installation wizard.
   - To run the installation program in text mode from a console, complete the following steps:
     - Change to the agent directory.
     - Enter the following command:

```
$ install.exe -i console
```

The screens are rendered in text and prompt you for the same information as the graphical installation.

> **Note:** After you enter the command, a process runs in the background for several seconds before the console installation program is displayed.

- To install the agent server silently, complete the following steps:
  - Change to the `Agent/responseFiles` directory.
  - Edit the `DistributedFileSystemCrawler.properties` template response file to provide information about your environment. To run the installation program, change to the agent directory, and then specify the name of the file that you edited.

    See the following example:

    ```
    $ install.exe -i silent -f responseFiles/DistributedFileSystemCrawler.properties
    ```

  If you copy a template file to another location to edit, specify the fully qualified path for the file when you run the installation program. If the response file path includes a space, enclose the path in double quotation marks (`"`). See the following example:

  ```
  $ install.exe -i silent -f "c:\My Documents\DistributedFileSystemCrawler.properties"
  ```

7. You must provide the following information during the installation process:
   - `hostname` : Enter or verify the fully-qualified hostname of the computer you are installing the agent server on.

     > ⚠ **Important:** You cannot specify an IPv6 address as the hostname of the server.

   - `username` : Enter the username of an account that can be used to authorize access to the agent server.

     If the username does not exist, select the checkbox to create the account.

     > ⚠ **Important:** To crawl a domain in a secure collection, the username must be an existing domain user with administration privileges for the Windows system to be crawled. To specify a domain user, use the format `<username>@<domain name>` .

   - `password` : Provide the password that is associated with the username.
8. **Optional**: If you want to change the default path and port settings, click **Advanced Options**.
   - You can change the paths for the installation directory and data directory.
   - The agent server uses three TCP/IP ports for authenticating connections to the server, transferring data between the file systems and Discovery, and monitoring the agent server. The default port numbers are `8397` and `8398`. If those values conflict with other port assignments in your system, change the port numbers.
9. On the summary page, review the options that you selected, and click **Install** to start installing the software.
10. **Optional**: If you want to secure traffic between the Windows Agent service and the crawler, enable TLS support.

    Copy the file named `tls.p12` from the decompressed directory to the root directory where the agent is installed. For example, the root directory might be `C:\Program Files\IBM\es\distributed\esadmin` .

    > **Note:** TLS support is available starting with the 4.7 release.

11. Restart your computer.

Configuring shared directories on the agent server

After the software is installed, you must set up shared network directories that the Windows File System agent can access. To define a new file system share, export a local or remote network directory.

> ⚠ **Important:** If you are replacing an agent that you installed with a release that is earlier than 4.6.0.0, skip this procedure. The replacement instructions explain how to reuse the file share that was defined previously.

1. Export a local directory from the server where the agent is installed:

   ```
   $ esagent --addshare <d:><\example>
   ```

Where `d:` represents the drive letter you want to use and where `\example` represents the path to the local directory.

2. Export a remote network directory that is accessible from the server where the agent is installed:

```
$ esagent --addshare <\\files.example.com\data>
```

Where `\\files.example.com\data` represents the hostname or IP address of the remote server or the path to the remote directory.

3. List shares that are defined on the server where the agent is installed:

```
$ esagent --lsshare
```

4. If you want to delete a share that is defined on the server where the agent is installed, you can use the following command:

```
$ esagent --rmshare \\files.example.com\data
```

Server status commands

After you install the agent server, you can enter commands to start, stop, and check the status of the server.

Stopping the agent server also stops the crawler. For example, if the crawler stops unexpectedly, you can close connections and release resources for that crawler.

- To start the server, enter the following command:

```
$ esagent start
```

- To stop the server, enter the following command:

```
$ esagent stop
```

- To get the status of the agent server, enter the following command:

```
$ esagent getStatus
```

The output of the `getStatus` command is an XML file with the following output:

```
<AgentStatus>
   <SpaceStatus>
      <SpaceId>012</SpaceId>
      <RootFolder>E:\\Projects\Analytics\\data\test1</RootFolder>
      <ConnectionNumber>9</ConnectionNumber>
      <StartTime>1244709336093</StartTime>
      <LastTime>1244709385843</LastTime>
      <IdlePeriod>219</IdlePeriod>
   </SpaceStatus>
   <SpaceStatus>
      <SpaceId>013</SpaceId>
      <RootFolder>E:\\Projects\Analytics\\data\test2</RootFolder>
      <ConnectionNumber>10</ConnectionNumber>
      <StartTime>1244709336093</StartTime>
      <LastTime>1244709385843</LastTime>
      <IdlePeriod>219</IdlePeriod>
   </SpaceStatus>
```

Connecting to a Windows File System data source

From your Discovery project, complete the following steps.

> **Note:** If you completed the prerequisite steps, return to the Windows File System data source collection that you started to create, and then skip to Step 4.

1. From the navigation pane, choose **Manage collections**.
2. Click **New collection**.
3. Click **Windows File System**, and then click **Next**.

4. Name the collection.
5. If the language of the documents that you want to crawl is not English, select the appropriate language.

   For a list of supported languages, see Language support.
6. **Optional**: Change the synchronization schedule.

   For more information, see Crawl schedule options.
7. In the *Enter your credentials* section, add values to the following fields. You provided these fields during the installation of the agent server, which was described in the Prerequisite steps section.

   Host

   The hostname of the remote Microsoft Windows server, for example `<hostname>.mydomain.com`.

   Username

   The username to connect the agent server. You use the username to connect Discovery to the shared network folders and crawl content.

   Password

   The password that is associated with the username.

   Agent Authentication Port

   The port to use for authentication. The default port value is `8397`.

   Port

   The port to use for transferring data. The default port value is `8398`.
8. In the *Specify what you want to crawl* section, enter the file path that you want to crawl in the **Path** field, and then click **Add**.

   The file path is case sensitive.

   Optionally, add more file paths.
9. **Optional**: Customize the types of files that are crawled.

   The crawler is configured automatically to exclude a list of file extensions for file types that can be unsafe to crawl. You can add more file extensions to the excluded filter list, or list only the file extensions for file types that you want to include in the crawl. Listing the types of files to include is even more secure.

   To change the file types that are crawled, in the *Extension filter* section, choose whether to use an Excluded or Included filter list. And then list the file extensions for the types of files you want to exclude or include.

   > **Note:** This configuration option was introduced with the 4.0.3 release.
10. **Optional**: Specify the character set of the data to crawl.

    The converter that is used by the crawler is configured automatically to detect the character set of the files before it converts them. However, you can choose to specify a different character encoding to use for the data conversion. To specify a character encoding, complete the following steps:

    - Set the **Automatic code page detection** switch to `Off`.
    - In the *Code page to use* field, specify the character encoding as a Java Charset value. For example, `UTF-8` or `UTF-16`. If you don't specify a character set, ISO-8859-1 is used.

    > **Note:** This configuration option was introduced with the 4.0.3 release.
11. **Optional**: If you want to enable document-level security, in the *Security* section, set the **Enable Document Level Security** switch to `On`.

    When you enable this option, your users can crawl and query content that they have access to. You must provide the details about the LDAP directory you want to use.

    LDAP server URL

    The LDAP server URL to connect to. For example, `ldap://<ldap_server>:<port>`.

    LDAP binding username

The username to use to bind to the directory service.

LDAP binding user password

The password that is associated with the binding username.

LDAP base DN

The starting point for searching user entries in LDAP. For example, `CN=Users,DC=example,DC=com`.

LDAP user filter

The user filter to search user entries in LDAP. If empty, the default value is `(userPrincipalName={0})`.

12. If you want the crawler to extract text from images in documents, expand *More processing settings*, and set **Apply optical character recognition (OCR)** to `On`.

    Note: When OCR is enabled and your documents contain images, processing takes longer. For more information, see Optical character recognition.

13. Click **Finish**.

The collection is created quickly. It takes more time for the data to be processed as it is added to the collection.

If you want to check the progress, go to the Activity page. From the navigation pane, click **Manage collections**, and then click to open the collection.

Enabling TLS for an existing collection

To ensure that all traffic that is sent between the Windows Agent service and the crawler is sent over the transport layer security (TLS) protocol, enable TLS support.

This capability is available starting with version 4.7. Do not complete this task until after you upgrade your service software to 4.7.

Important: After you enable TLS for the Windows Agent service, any existing collections in deployments with earlier versions of Discovery will not be able to connect to this Windows Agent service.

To add TLS support to an existing collection, complete the following steps:

1. Open the *Processing settings* page for the existing Window File System collection.
2. Install the latest version of the agent.

   Complete the steps in the Installing the agent procedure, starting with Step 4 and including the optional step to enable TLS support.

   Important: Do not complete the last step that asks you to restart your computer.

3. Find and open the `as.cfg` file in a text editor, and then add the following lines to the file:

   ```
   agent_key_store=%ES_AGENT_NODE_ROOT%\tls.p12
   agent_key_store_password=changeit
   ```

   where `%ES_AGENT_NODE_ROOT%` is the root directory for the Windows Agent server. For example:

   ```
   agent_key_store="C:\Program Files\IBM\es\distributed\esadmin\tls.p12"
   agent_key_store_password=changeit
   ```

4. Restart the Windows Agent service by using the following commands:

   ```
   esagent stop
   esagent start
   ```

Troubleshooting ingestion

Learn about solutions and workarounds to warnings or errors that you might encounter when you add data to a collection.

Note: This information applies both to managed and installed instances of Discovery. For more troubleshooting tips for installed deployments only, see Troubleshooting IBM Watson® Discovery Cartridge for IBM Cloud Pak® for Data deployments.

Documents fail to index with a mapper parsing exception error

Discovery provides a rich set of query and aggregation functions for all supported field types such as `nested`, `string`, `date`, `long`, `integer`, `short`, `byte`, `double`, `float`, `boolean`, and `binary`. To support the functions and remain schema-less, collections in Discovery automatically detect a field data type when the field's data is first parsed during an add or update document process.

Consider that the following document is first ingested to a collection.

```
{
"foo": "lorem ipsum",
"bar": 12345,
"baz": "2024-01-01"
}
```

The automatically detected field data types for the first document are shown in the following response of the `List fields` method. For more information, see List fields in the API reference.

```
{
      "fields": [
            {
               "field": "foo",
               "collection_id": "6537819f-8a3d-de55-0000-018d199c9c1e",
               "type": "string"
            },
            {
               "field": "bar",
               "collection_id": "6537819f-8a3d-de55-0000-018d199c9c1e",
               "type": "double"
            },
            {
               "field": "baz",
               "collection_id": "6537819f-8a3d-de55-0000-018d199c9c1e",
               "type": "date"
            },
            ...
      ]
}
```

Now, when subsequent documents are added to the collection, they must have field data that is supported by each automatically detected field type from the first document, or must not have that field at all. If a subsequent document contains field data that is not supported by the respective field type, then Discovery fails to index the entire document and a `mapper_parsing_exception` error occurs.

For example, consider that the following second document is next ingested to the same collection:

```
{
"foo": "dolor sit amet",
"bar": 67890,
"baz": "consectetur adipiscing elit"
}
```

Ingesting the second document fails because the `baz` field contains field data (`consectetur adipiscing elit`) that cannot be parsed as `date`, which is the automatically detected field type for the `baz` field from the first document ingestion process. The unsupported field data for the `baz` field results in the following `mapper parsing exception` error.

```
{
   "severity": "error",
   "created": "2024-01-17T23:05:30.968Z",
   "description": "Failed to index. type=\"mapper_parsing_exception\", reason=\"failed to parse field [baz] of type [date] in document with id
'b326428e0ce9a2e829327d393b14d76f'. \"",
   "step": "indexing",
   "document_id": "b326428e0ce9a2e829327d393b14d76f",
   "customer_id": "",
   "notice_id": "index_failed_elastic_return_error"
}
```

To resolve this error, plan to ingest documents in a sequence that sets the automatically detected field types to more permissive field types such as `string`.

To add both the first document and second document in the examples to the same collection, you must set the `baz` field data type to the more permissive `string` field type. To do so, you can reverse the order in which the documents are ingested to the collection. So, when the collection first parses the `baz` field, the field type is automatically detected as `string`. In general, the `string` field type accepts any data format.

Alternatively, navigate to the **Manage fields** page. To access the **Manage fields** page, click the **Manage collections** icon in the navigation panel, open the collection, and then click the **Manage fields** tab. You can change the field type of top-level fields to more permissive types such as `string` in the **Manage fields** page and then reprocess the collection. If the field type is already selected as `string`, choose a different field type and select `string` again, then click **Apply changes and reprocess**.

For example, the following image shows the field type for the `baz` field.

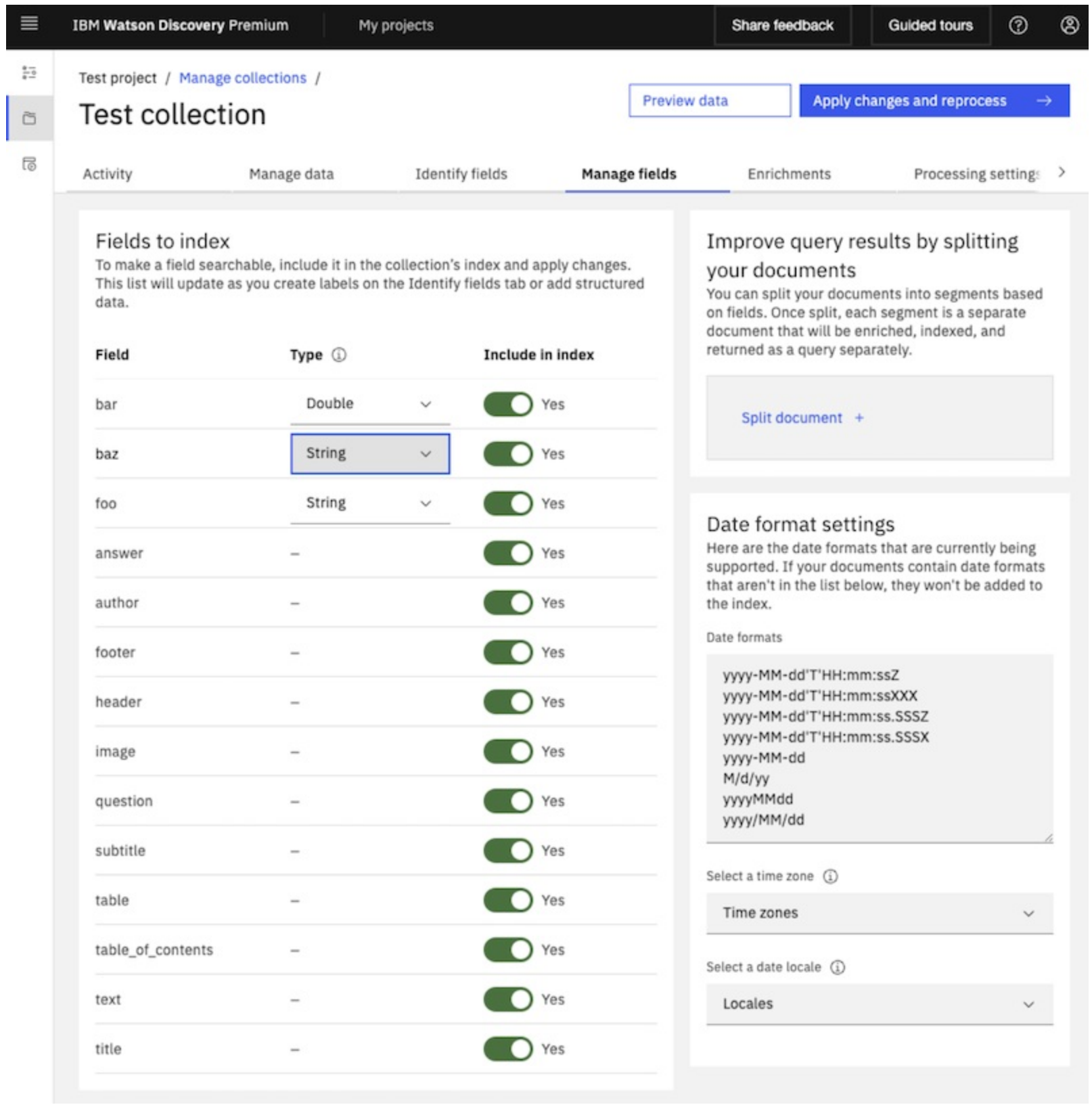

Figure 1. Manage fields page

After applying changes and allowing the collection time to reprocess, you can ingest the second document in the example successfully to the same collection. For example, the following response from the `Get document details` method shows that the second document is ingested successfully. For more information, see Get document details in the API reference.

```
{
   "document_id":  "b326428e0ce9a2e829327d393b14d76f",
   "created":  "2024-01-17T23:04:00.411Z",
   "updated":  "2024-01-18T00:00:36.158Z",
   "status":  "available",
   "notices":  [],
   "children":  {
      "count":  0,
      "have_notices":  false
   },
   "filename":  "second.json",
   "file_type":  "json",
   "sha256":  "ba280879c7c30885478563ee14e0fbb23186eaeecdf5d554c7f50efd9bab4a35"
}
```

Unable to process one or more documents

This notification is displayed in the page header when a processing delay of any kind occurs in any project across the entire service instance. If the message is displayed while you are adding data to a collection, you can ignore it. If any problems occur that are related to the creation of your collection, a message is displayed in the *Activity* page for the collection. Check there for any pertinent messages.

This document exceeds the 1MB limit for non-HTML fields

> The `html` field in the document index stores structural information about the document. If you add a single document with complex tables, images, or other objects that need to be represented in HTML, you might hit the size limit for this field. To work around this issue, consider breaking the source file up into 2 or more smaller files, and then add the files to the same collection separately so that you can apply enrichments and search them together.

Microsoft document troubleshooting tips

Failed to prepare document for SDU processing

> Some DOC, PPT, and XLS files that use older features which are no longer supported by Microsoft Office can cause ingestion issues. If you encounter this issue, open the file in a more recent version of Microsoft Office and convert the file to the DOCX, PPTX, or XLSX format respectively, and then upload the DOCX, PPTX, or XLSX file.

Line breaks are inserted randomly

> When some files in Microsoft Office format are added to a collection, line breaks are inserted seemingly at random to the text that is stored in the `html` field in the collection's index. The unexpected line breaks can impact the efficiency of enrichments, such as custom rule recognition.
>
> Cause: As part of their ingestion into Discovery, such files are converted from Office format to PDF format. When the conversion happens, textual content is sometimes lost due to the nature of a PDF file. While the new lines appear to be added at random, they typically get inserted in areas where text wraps in the original document, such as in narrow text boxes or to accommodate other inline elements, such as images or diagrams.
>
> Solution: To avoid new line insertions, increase the width of text boxes in the original document. If the original document has a section where text wraps to accommodate an inline element, such as an image, move the image so that it is situated in its own section and the nearby text doesn't need to wrap around it. To test whether your fixes address the issue, you can convert the original file to a PDF file to check for unexpected carriage returns in the text.

After applying a pretrained Smart Document Understanding model to a PPT file, table boundaries are not recognized properly

> During the conversion process, text that is extracted from the table is confused with text that is outside the table in some PPT pages. This issue is more likely to occur in tables with a lot of text and that have footnotes displayed just outside the table border. If you encounter this issue, export the PPT file as a PDF file, and then upload the PDF file instead. Apply a user-trained Smart Document Understanding (SDU) model to the document, and then use the SDU tool to identify the tables in the document. The resulting model handles table boundaries properly and can extract text from the tables cleanly.

PDF file troubleshooting tips

Failed to parse document due to invalid encoding

> Enable OCR for the file.

Enrichment troubleshooting tips

Table Understanding: *n* input tables excluded by enrichment

> If tables in a document have inconsistent column and row spans or are too large for the system to process completely, the table understanding enrichment is not applied to them. Information from such tables cannot be returned in search results. If you want the table understanding enrichment to be applied to a table that was skipped, consider editing the table. Change a table with inconsistent column and row spans to have a simpler table format. Split a large table into many smaller tables.
>
> To find the table where the enrichment was not applied, check the warning message. It lists the character offsets where the table begins and ends in the HTML representation of the document. To see the full warning message and get the document ID, click **View all**, and then make a note of the document ID. From the *Improve and customize* page, submit an empty search query to return all of the indexed documents. Look for the document ID. (You can change the search result settings to show the document ID as the result title.) Click the **View passage in document** link for your document, and then click **Open advanced view**. Choose to view the document as JSON and then look for the `html` field. Copy and paste the HTML representation of the document into a text editor. Look for the character offsets that were listed in the original warning message to find the table.

Managing data collections

After the processing of a new data collection is finished, you can see a summary of the settings that are applied to your collection from the *Manage collections* page.

For more information about how to create a collection, see Creating collections.

Managing data

Note: The *Manage data* page is available in installed deployments starting with the 4.6.5 release of IBM Cloud Pak for Data.

After you create a collection and the documents in the collection are indexed, you can see a list of the documents from the *Manage data* page.

1. Open the *Manage collections* page.
2. Click to open the collection that you want to change.
3. Click the **Manage data** tab.

   A list of the documents in the collection is displayed.

   IBM Cloud To preview a document in the collection in the advanced document view, click **Preview data**.
4. **Optional**: You can change the information that is displayed.

   To change the fields that are shown in the view, click the *Settings* icon at the start of the view. Choose a different field for the first and second columns, and then click **Apply**.

   For example, you can change the fields in the view to accomplish the following goals:

   - Get the document ID for a document that you want to work with by using the API.
   - Find the parent document for a document. Some file types, such as CSV or JSON files, generate subdocuments when they are added to a collection, for example. And splitting a document turns one document into multiple document segments.
   - Retrieve the original file name for a document.
   - Find out how many pages are in a document.

   Note: The custom settings that you apply are not retained. The default field settings are shown the next time that you access the page.
5. **Optional**: You can delete a document from the collection from this page. For more information, see Excluding content from query results.

Changing how a data source is processed

You can change settings that were applied to a collection when it was created. You might want to change the schedule at which an external data source is crawled, for example.

To change how a data source is processed, complete the following steps:

1. Open the *Manage collections* page.
2. Click to open the collection that you want to change.
3. Click the **Processing settings** tab.
4. Make any changes that you want to make to the processing settings.

   For example, you might want to enable or disable optical character recognition (OCR), which is a feature that extracts text from images. For more information, see Optical character recognition.

   For more information about changing data synchronization schedules, see Crawl schedule options.

   Other setting options differ by data source type.
5. Click **Apply changes and reprocess**.

Finding where a collection is used

To find out whether a collection is being shared, open the *My Projects* page, and then complete the appropriate step for your deployment:

- IBM Cloud Pak for Data Click **Collection usage and sharing**.
- IBM Cloud Click **Data usage and GDPR**, and then review the **Collection usage** page.

Collections can be associated with a single project, shared by two or more projects, or not associated with any project.

Reusing data from a collection

When you share collections across multiple projects, the following resources are shared:

- The processed data
- Configured connector

If you make any of the following changes to a shared collection, the changes are applied to the collection in every project where it is shared:

- Changing the Optical Character Recognition (OCR) setting
- Annotating fields or adding fields by using Smart Document Understanding
- Enabling or disabling fields
- Changing the setting for document splitting
- Changing any of the connector settings

> ⚠ **Important:** Enrichments and improvement tool settings are not included when a collection is shared because they are set at the project level.

For more information about the other tabs, see the following topics:

- **GDPR data label** IBM Cloud: For more information about GDPR and labeling data, see European Union General Data Protection Regulation (GDPR) .
- **API usage** IBM Cloud Pak for Data For more information about monitoring Analyze API usage, see Monitoring usage.

Deleting collections

Find out whether a collection is being used anywhere before you delete it from the *Collection usage* page. Unshared collections can be deleted directly from this page.

- To delete a single collection from a project, open the *Manage collections* page from the navigation panel, find the collection tile, and then click the delete icon.

  Decide whether to keep the underlying data and configuration settings. If you choose to keep the data, you can find the collection in the unshared list on the *Collection usage* page. You might need to wait a few minutes before the collection is displayed.

  Click **Delete from project**.

- IBM Cloud Pak for Data To delete all of the collections in your environment, select the Environment details icon , and then choose **Delete environment**.

  > **Tip:** *Environment* refers to the Discovery instance that you provisioned in IBM Cloud Pak for Data.

You cannot delete the *Sample Project* collection.

How your data is processed

When you connect to a data source, Discovery *processes* the information from the data source to create a *collection*.

The goal of processing a data source is to identify meaningful information and tag it as it is added to the collection so it is easier to find and retrieve the information later.

The processing that is applied to all data sources includes the following steps:

- Identify individual documents in the data source
- Find fields in the documents
- Index the fields

You can see a list of the fields that were indexed from the *Manage fields* page.

1. Go to the *Manage collections* page, and then choose the collection.

   Make sure that the processing of the collection is finished first. The Activity page shows the processing status.

2. Click the *Manage fields* tab.

The fields that are shown can differ based on your data. However, one subset of fields is always listed. These fields, with names such as `footer` and `header`, are derived from the Smart Document Understanding (SDU) tool, and are listed even when you don't explicitly apply an SDU model to the collection. (For the full list of SDU-generated fields, see Available fields.) Only the fields with a data type specified are stored in the collection's index.

One of the SDU-generated fields that is stored in the index is the `text` field. The `text` field typically contains the main body of text from the original document. Most of the content that is returned in search results that you submit from the *Improve and customize* page originates from this one field. How to parse and return only relevant chunks of information from this field is determined by the query result configuration that is used by the project. For more information, see Previewing the default query results.

More processing adds more fields. And more processing is applied automatically depending on the project type. When processes run on documents in a collection, extra fields are added to store information that is associated with the process. For example, when the built-in Entities enrichment is applied to a collection, it starts a process that adds fields with names that begin with `enriched_{field_name}.entities` to the documents in the collection.

- For more information about the enrichments that are applied by default, see Default project settings.

How fields are handled

For most unstructured file types, the bulk of the content from the file is added to a field named `text`. For file types that have an inherent data structure, such JSON files, names from the source file are used to name the fields in which the content is stored. When you upload files of this type, be aware of some naming limitations that exist for fields.

The following field names have special meaning. If possible, do not use these names in your structured source files.

- `document_id`
- `highlight`
- `html`
- `metadata`
- `parent_document_id`
- `result_metadata`
- `score`
- `spans`

Avoid field names that meet the following conditions. Field names with these restricted characters are not queried.

- Start with the characters `_`, `+`, and `-`. For example, `+extracted-content`.
- Contain the characters `.`, `,`, `#`, `?`, `(`, `)`, or `:` or spaces. For example, `extracted content` or `new:extracted-content`.
- End with numbers, for example, `extracted-content2`.

Note: To process documents in Discovery, all documents in a collection must have the same data-type for a particular field. When a data-type of a particular field varies among documents, the field indexing process fails and a failed to index error message is displayed in the *Warnings and errors at a glance* section of the *Activity* page for the collection.

HTML fields

The `html` field in the document index stores structural information about the document.

- If you use the Smart Document Understanding tool to annotate a collection, the document representation is indexed in the `html` field.
- If you use the Smart Document Understanding tool to apply a pretrained model to a collection, the document representation is indexed both in the `html` field and `text` field.
- The `html` field has a size limit. For more information, see Field limits.

Note about enhancing data:

- If you want to apply an enrichment that can understand the tables in a document, the document must contain an `html` field.

How dates are handled

Dates are captured in different ways by different file types.

Unstructured files

The best way to capture date information from the body of a document with unstructured data is to use a natural language processing model enrichment. For example, the prebuilt Entities enrichment recognizes dates and annotates them in the `text` field (or other body fields with the `String` data type). In a document where the enrichment is applied, you can find dates by looking for fields that are labeled as `enriched_{fieldname}.entities.type` = `Date`.

Dates from metadata date fields, such as `extracted_metadata.publicationdate`, are stored in the index as dates as long as the date format matches one of the supported date data type formats. You can't see nested fields from the *Manage fields* page. And when you view a search result as JSON, date field values are displayed as string values because the JSON editor shows the date as a string. However, values from date fields behave like dates. You can use greater than (`>`) or less than (`<`) operators with such fields in Discovery Query Language queries, for example.

Structured files

Structure files that you import, such as CSV or JSON files, might contain date fields that you want to store as date data types. Discovery can recognize many date formats. However, you might need to add a format to the list. For more information, see *Date format settings*.

Date format settings

If your documents have a root-level field with date information in it, you can set the field to be a `Date` data type field in the index.

Discovery recognizes the following date formats automatically:

```
yyyy-MM-dd'T'HH:mm:ssZ
yyyy-MM-dd'T'HH:mm:ssXXX
yyyy-MM-dd'T'HH:mm:ss.SSSZ
yyyy-MM-dd'T'HH:mm:ss.SSSX
yyyy-MM-dd
M/d/yy
yyyyMMdd
yyyy/MM/dd
```

If you store dates in other formats, you can add the format to the list of supported formats.

To add more date formats, complete the following steps:

1. From the *Manage fields* page for the collection, add a format as a new line in the **Date formats** field.

   Specify a date format that is supported by the Java SimpleDateFormat class.

   For example, if your records store only year values for dates, add `yyyy` to the supported date formats list. You can then set the data type for the field that contains a year value to *Date*, and reprocess your collection. As a result, an occurrence of `2019` in the date field is stored as `2019-01-01T05:00:00Z` in the index.

   When you add a date format, you must specify an associated time zone for the date.

2. Specify a time zone.

3. Optionally, select a date locale.

   The locale that you choose is used to parse a string value that represents the date for the date-type data set fields. For example, by using the `EEE, MM dd, yyyy` format, the **English (United States)** locale can parse the string value of `"Wednesday, 07 01, 2020"`, and the **Japanese (Japan)** locale can parse the same string value of `"水曜日, 07 01, 2020"`.

4. If you already imported documents with dates in formats that were not recognized, reprocess the documents.

Discovery cannot store a date that is mentioned within a text field as a *Date* field in the index. However, you can use an enrichment such as the *Entities* enrichment to identify dates that are mentioned in text.

How file types are handled

When you upload a document, data in the file is indexed. Different files types are handled differently by Discovery.

- CSV files
- HTML files
- JSON files

CSV files

Notes about adding data:

- Each line that is defined in the CSV file is added to the index as a separate document, each with the same `parent_document_id`.

  The child documents typically have a document ID with the syntax `{parent-ID}_n` where *{parent-ID}* is the document ID of the original file that was added and *n* is a sequential number. For example, if you upload a CSV file with 5 rows, then five documents are added to the collection with document IDs such as `f5214225c1e03e25190ffcdfad8e84ff_0` through `f5214225c1e03e25190ffcdfad8e84ff_4`.

- You cannot enable the Optical Character Recognition (OCR) feature for CSV files.

- If the CSV file has headers, the header names are used to name the fields in which the content from the corresponding column is stored. Do not use names that have special meaning in Discovery. Be sure that the field names conform to the naming rules, such as having no spaces and no appended numbers. For example, you can rename the `start date` header to `start_date` and `label1` to `label-one` before you add the file. For more information, see How fields are handled.

- When a CSV file header name contains restricted characters, the document converter automatically removes the restricted characters from the field name when it adds the resulting field to the index.

Note about enhancing data:

- You cannot apply prebuilt or user-trained Smart Document Understanding models to CSV files.

HTML files

If you upload an HTML file or crawl a data source with HTML files, such as a website, an `html` field is generated along with the `text` field. For more information, see HTML fields.

JSON files

Notes about adding data:

- Object names from the source JSON file are used to name the fields in which the content is stored. Do not use names that have special meaning in Discovery. Be sure that the names conform to the naming rules, such as having no spaces and no appended numbers. For example, you can rename the `updated on` object to `updated_on` and `answer2` to `answer-two` before you add the file. For more information, see How fields are handled.

- If a root-level field is an array but contains no items, the field is omitted from the index.

- If a root-level field is an array and contains only one item, the array is indexed as the data type of the one item. For example, a string array with one string is indexed as a string.

- If a nested field contains an array, even if the array has only one value, it is indexed as an array.

- If a root-level field is an array and contains more than one item, the data is indexed as an array.

- If you copy JSON that is generated by Discovery and then upload it as a JSON file, remove these system-generated fields from the file first: `document_id`, `parent_document_id`, `filename`, and `title`.

- You cannot enable the Optical Character Recognition (OCR) feature for JSON files.

- If your source document has a field with the name `document_id`, the field is skipped and not added to the index in the collection.

> **Note:** How the `document_id` field in a JSON file is handled changed with the `2023-03-31` version update of the API. Before the update, when you uploaded a JSON file from the product user interface or used the API to add it with the *Add document* method, the value in the `document_id` field from the file was shown as the `document_id` value in query results. However, a different document ID was assigned to it and stored in the `parent_document_id` field. The assigned document ID is what was returned when you called the *List documents* method and is what had to be used as the `document_id` in the endpoint URL for a *Delete document* method request. When you used the *Update document* method to assign a new `document_id`, the original ID continued to be returned in query results. However, the assigned ID had to be used to delete the document. If you have an application that relies on the previous behavior, you can specify a version number earlier than 2023-03-31, such as `2020-08-30`, in your API calls.

Notes about enhancing data:

- You cannot apply prebuilt or user-trained Smart Document Understanding models to JSON files.

- When you apply an enrichment to a field from the JSON file, the field data type is converted to an array. The field is converted to an array even if it

contains a single value. For example, "field1": "Discovery" becomes "field1": ["Discovery"].

- Only the first 50,000 characters of a custom field from a JSON file are enriched.
- In project types where the *Part of Speech* (POS) enrichment is applied automatically, the enrichment is applied to the field that contains the bulk of the file content in the first JSON file that is added to the collection. This field is determined by the following rules:
  - If a field is named `text`, the POS enrichment is applied to it.
  - The field with the longest string value and highest number of distinct values is chosen.
  - If more than one field meets the previous condition, one of the fields is chosen at random.
- If you want to apply an enrichment to a nested field, you must create a Content Mining project, and then apply the enrichment to the field. If you want to use a project type other than Content Mining, you can reuse the collection that you created with the Content Mining project type elsewhere. For more information, see Applying enrichments.

Note: You can specify the `normalizations` and `conversions` objects in the Update a collection method of the API to move or merge JSON fields.

How passages are derived

Discovery uses sophisticated algorithms to determine the best passages of text from all of the documents that are returned by a query. Passages are returned per document by default. They are displayed as a section within each document query result and are ordered by passage relevance.

Discovery uses sentence boundary detection to pick a passage that includes a full sentence. It searches for passages that have an approximate length of 200 characters, then looks at chunks of content that are twice that length to find passages that contain full sentences. Sentence boundary detection works for all supported languages and uses language-specific logic.

For all project types except *Conversational Search*, you can change how the passages are displayed in the search results from the **Customize display > Search results** page. For example, you can configure the number of passages that are shown per document and the maximum character size per passage.

Querying your data from the UI

Previewing your query results

See the types of query results that are returned automatically and learn about how they are derived.

When a document is ingested, the text is extracted and indexed in the `text` field. To return only the subset of information that is relevant to the query, Discovery returns *passages* from the `text` field. For more information about passages, see How passages are derived.

When you enter a query from the product user interface, the UI submits the text as a natural language query. For more information about how to query your data programmatically, see Query API.

To query your data from the product user interface, complete the following steps:

1. From the navigation panel, open the *Improve and customize* page.
2. Take the appropriate next steps for your project type.
    - Document Retrieval
    - Conversational Search
    - Document Retrieval for Contracts
    - Content Mining

Document Retrieval

1. Do one of the following things:
    - Click **Run search** for one of the keywords that Discovery calculated to have special meaning in your collection.
    - Submit your own phrase or keyword from the search bar.

    You can see that the query results that are returned consist of passages.

    Entities that are recognized in your documents (based on the Entities enrichment that is applied to the project by default) are displayed as facets by which you can filter the query results.
2. To explore a query result in more detail, click **View passage in document**.
3. Click **Open advanced view** to explore the entity mentions that are recognized by Discovery.

Excerpt unavailable

Passages displayed in search results are extracted from the content that is indexed in the `title` and `text` fields of the documents. If the content is indexed in other fields, the search displays the `Excerpt unavailable` message.

Content is indexed in other fields in the following scenarios:

- Your collection contains structured files, such as JSON or CSV files. When you ingest structured files, content is stored in custom fields with names taken from the original object names (JSON) or column headers (CSV).
- You applied a Smart Document Understanding model that moves content from the `text` field into new fields, such as `section` or `results`, based on the document's structure.

To improve the search results, first choose how you want to extract content from the documents. If your documents have targeted fields with succinct content in them, such as `answer` fields for an FAQ use case, configure the search to return those specific fields. If your documents have custom fields with lots of content in them, such as `chapter` fields, configure the search to find the best passages from the custom fields.

To customize the results, complete the following steps:

1. From the *Improvement tools* panel, expand **Customize display**.
2. Click **Search results**.
3. In the *Select source of result content* option, do one of the following things:
    - Select **Passages**, and then specify one or more fields from which to extract passages.
    - Select **Field**, and then choose one or more fields to return.

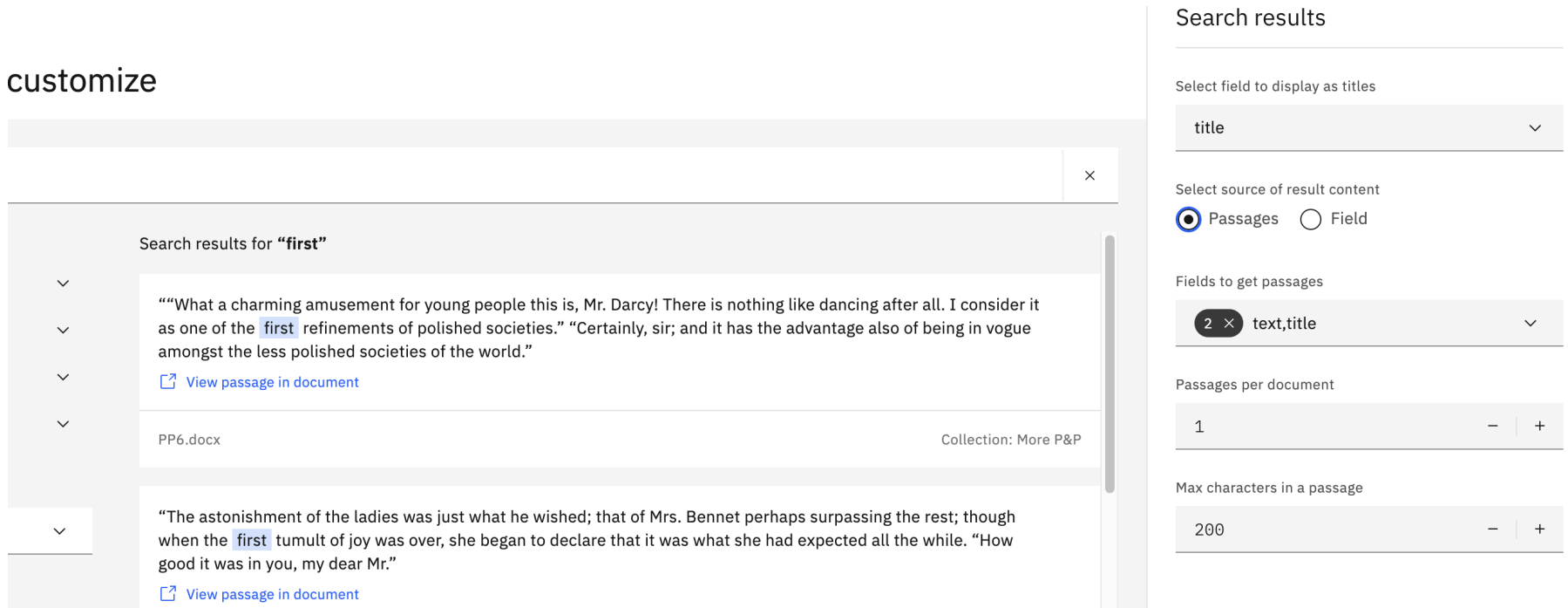

Figure 1. Search results dialog

4. Click **Apply**.

Conversational Search

A single search field is displayed that mimics the user interface of a virtual assistant.

1. Submit a phrase or keyword.

The query results are returned as passages by default. You can configure the search to return a specific field instead.

If you want to investigate the results a bit more, you might want to use a different project type. For more information, see Improving results for a chatbot.

Document Retrieval for Contracts

Contract-related elements that are recognized in your collection are displayed.

1. Do one of the following things:
    - Filter the documents by one of the highlighted elements or by entities that are recognized in your documents (based on the Entities enrichment that is applied to the project by default).
    - To view the contract elements in more detail, click a document result to open it. Open the *Contract Data* tab.

For more information about the elements, see Understanding contracts.

Content Mining

Facets based on the *Part of Speech* enrichment are shown.

1. To analyze your data, open the Content Mining application. Click **Launch application**.

For more information, see Analyzing your data.

What to do next

- For more information about how to enrich your documents so that you can find key information, see Choosing enrichments.
- To explore ways to improve the query results, see Improving your query results.

Improving your query results

Learn about actions you can take to improve the quality of your query results.

You can use the tools that are built in to Discovery to make improvements.

Results include more than exact matches

Unlike some other search applications, adding quotation marks to a phrase that you submit does *not* return only exact matches. Queries that are submitted from the product user interface are natural language queries. When quoted text is submitted in a natural language query, the phrase is used to boost result scores. However, results are not limited to documents that contain the entire phrase.

If you want more control over how queries are handled, you must use the query API. For more information about the `phrase` operator of the query API, see Query operators.

A short query returns irrelevant results

It might be that your query contains too many stop words and not enough distinct terms to trigger a meaningful search. When you submit a query, the query text is analyzed and optimized before it is submitted to the project. One of the changes that occurs is the removal of any stop words from the text. A *stop word* is a word that is considered to be not useful in distinguishing the semantic meaning of the content. Examples of stop words include terms such as `and`, `the`, and `about`. Discovery defines a list of stop words that it ignores automatically both when the data is indexed and when it is searched. When you submit a query that contains mostly or only stop words, such as `About us`, it is equivalent to submitting an empty query.

Note: Although *us* is not included in the stop words list, it is lemmatized to *we*, which is listed as a stop word.

You can edit the stop words that are used by your collection. However, you can only augment the stop words list; you cannot remove stop words. And the stop words that you define are used only at query time. They do not affect the stop word list that is used by Discovery when data is added to a collection and the index is created.

For more information, see Identifying words to ignore.

Results have too much text

If the source document is large, consider splitting the document into smaller chunks.

To do so, you can create a Smart Document Understanding user-trained model. Find content in the document that can be used to consistently break your document into subsections. For example, maybe your document has chapters or subtitles. You can label the chapters with a custom label named `chapter`. After you teach the model to recognize the `chapter` content type, apply the model to your entire collection. For more information, see Using Smart Document Understanding.

You can then split the document by the `chapter` field to create many subdocuments segmented by chapter. For more information, see Split documents to make query results more succinct.

Information from tables is not found

The table understanding enrichment must be applied to your collection for information from tables to be searchable. The table understanding enrichment is applied to collections automatically in some situations. If it isn't and your collection has an HTML field in its index, you can apply the *table understanding* enrichment yourself.

For more information, see Understanding tables.

Information from diagrams is not represented

Text from diagrams and other images is not captured unless you enable the optical character recognition (OCR) setting for the collection. You can apply the setting to a collection after its initial creation. For more information, see Managing data collections.

Search does not recognize significant terms

If the results suggest that keywords, common nouns, or domain-specific terms in the query are not being recognized as significant, enrich your collection.

Use Watson Natural Language Understanding to find and tag terms that are generally understood to have special meaning, such as locations or company names. For more information, see Applying prebuilt enrichments.

Teach Discovery about terms and patterns that have special meaning to your use case. For more information, see Adding domain-specific resources.

Default facets aren't useful

You can add facets that categorize documents based on data from enrichments that you apply to a collection. For example, you might want to show facets based on keywords or dictionary categories. For more information, see Facets.

Explore other search features

When you test your project from the Discovery user interface, you submit a natural language query. Search features are available that you can enable to influence how the natural language query search is done. And Discovery Query Language search is another type of search that you can leverage by using the API. If the initial results don't meet your needs, experiment with another search method.

- Discovery Query Language (DQL) search: A search mechanism that accepts more complex queries. You must use the query API to submit DQL queries.

  For example, you can search for specific values in fields that are generated by enrichments that are applied to a collection.

- Natural language query is the type of search that is triggered from the *Improve and customize* page.

For more information about the Query API, see Query API overview.

Adding facets

Add more facets that you can use to filter your data.

When you apply custom enrichments to your collection, annotations are added to its documents. The annotations feed into new facets that you can use to sort your data.

The following table describes the types of facets that you can create from annotations.

Information to recognize	Annotator type
Commonly understood terms, such as organization or people names.	Built-in Natural Language Processing models
Phrases that express an opinion and evaluate whether the opinion is positive or negative.	Phrase sentiment
Alternative words that share a meaning with terms in a finite list.	Dictionary
Terms that match a syntactical pattern	Regular expression
Custom terms by the context in which they are used.	Machine learning model
Documents that fit into categories that you define.	Document classifier

Custom facet types

Grouping facets

To organize your facets, you can group them in folders.

Grouping facets does not combine the data from the facets. It merely makes the facets easier to find because they are organized in named folders.

To associate facets such that you can combine data from multiple facets, the facets must have a facet and subfacet relationship. Such hierarchical relationships must be defined at the time that the facet enrichment or annotation is created and applied to the collection.

To group facets, complete the following steps:

1. From the initial search page, submit a search.
2. From the *Facet analysis* pane, click the *Edit* icon.
3. Name the group, and then select the facets that you want to group together.

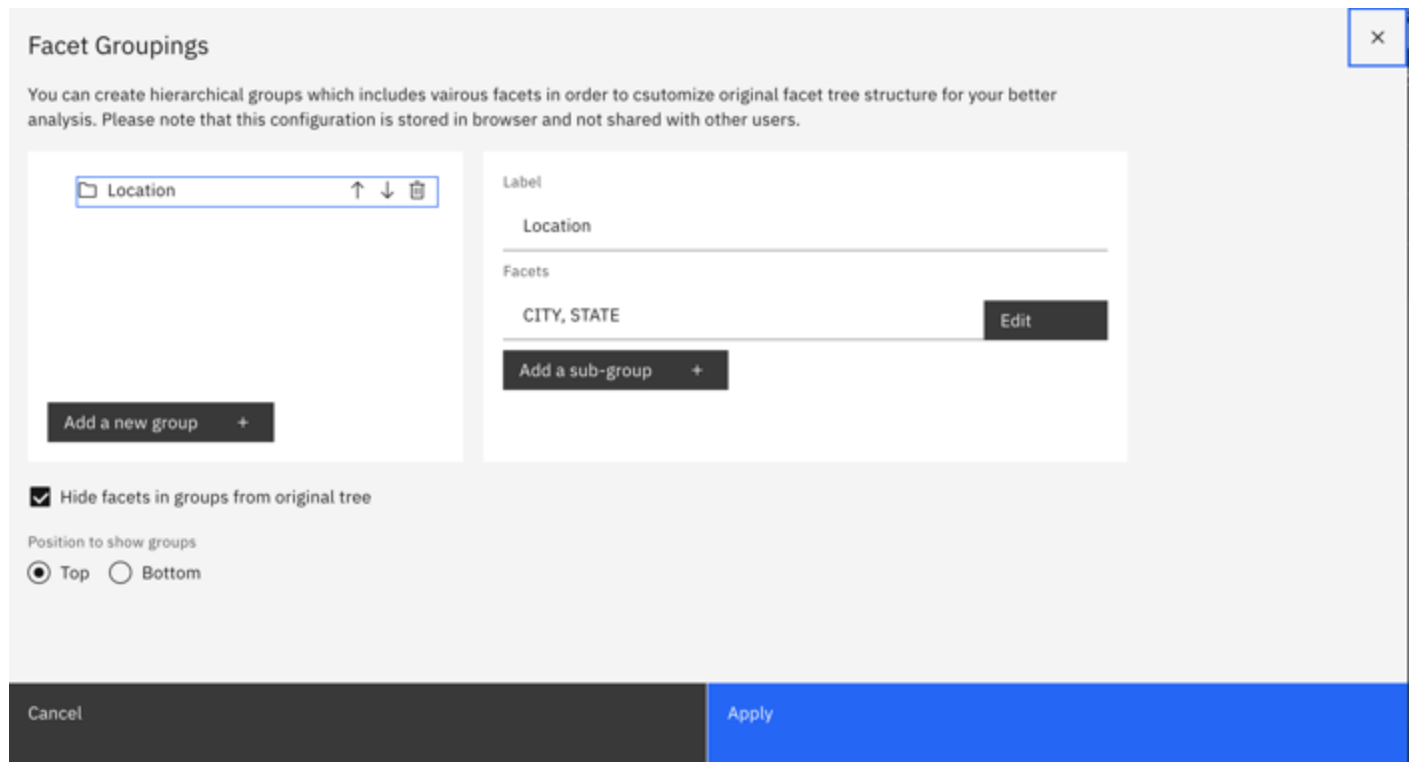

Figure 1. Facet grouping dialog

4. Click **Apply**.
5. The facets that you grouped are now available from a folder with the group name.

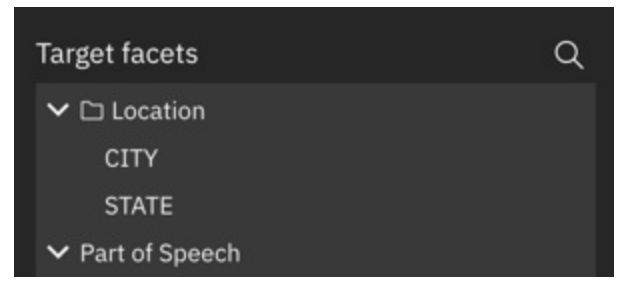

Figure 2. Facet folder from the facet list

Customizing the search bar

Control how customers interact with the search bar.

Decide whether the search bar can interact with customer query submissions in the following ways:

- Propose alternative search terms when a misspelling is detected
- Propose better query wording with type-ahead

Note: These search bar customizations are available for all project types except *Conversational Search*. Similar features, such as autocorrection, can be configured for the chat widget where you deploy the project.

To customize search bar behavior, complete the following steps:

1. From the navigation pane, open the **Improve and customize** page.
2. Expand **Customize display** from the *Improvement tools* pane, and then click **Search bar**.
3. Turn the following features on or off by setting the associated switcher:

Autocompletion

As the customer types a word as part of a query into the search bar, completed words are displayed as suggestions. The user can click a suggestion to add it to the query. The suggested words are based on terms from the project documents. Suggestions are not based on terms from the user's search history and the project does not learn from the user's choices. This setting is enabled by default.

Spelling suggestions

Recognizes words that are misspelled in the customer query. After the query is submitted, a `Did you mean:` link is displayed that shows a corrected version of the original query. The customer can click the corrected query to submit it. This setting is disabled by default.

Expanding the meaning of queries

You can improve the quality of search results by expanding the meaning of the queries that are submitted by customers.

To expand the scope of a query beyond exact matches, add a synonyms list to your collection. When synonyms are defined, the customer does not need to submit an exact phrase or keyword that your project is trained to understand. Even variations of the term are recognized and used to find the best results. For example, you can expand a query for `ibm` to include `international business machines` and `big blue`. Query expansion terms are typically synonyms, antonyms, or common misspellings for terms.

Note: Synonyms that you add to improve the search results function differently from synonyms that you add to a dictionary. Dictionary synonyms are recognized and tagged at the time that a document is ingested. The synonyms that you define are recognized and tagged as occurrences of the associated dictionary term, so that they can be retrieved later by search. For more information about adding synonyms that are recognized when documents are processed, see Dictionaries.

You can define two types of expansions:

Bidirectional

Each entry in the `expanded_terms` list expands to include all expanded terms. For example, a query for `ibm` expands to `ibm OR international business machines OR big blue`.

Bidirectional example:

```
{
   "expansions":  [
      {
```

```
        "expanded_terms":  [
           "ibm",
           "international business machines",
           "big blue"
        ]
     }
  ]
}
```

Unidirectional

The `input_terms` in the query is replaced by the `expanded_terms`. For example, a query for `banana` is converted to `plantain OR fruit` and does not contain the original term, `banana`. If you want an input term to be included in the query, then repeat the input term in the expanded terms list.

Unidirectional example:

```
{
   "expansions":  [
      {
         "input_terms":  [
            "banana"
         ],
         "expanded_terms":  [
            "plantain",
            "fruit"
         ]
      },
      {
         "input_terms":  [
            "car"
         ],
         "expanded_terms":  [
            "car",
            "automobile",
            "vehicle"
         ]
      }
   ]
}
```

To enable query expansion, complete the following steps:

1. Create a synonyms list file. The file must be a JSON file with the `json` file extension.

   Follow these guidelines:

   - Specify the `input_terms` and `expanded_terms` in lowercase. Lowercase terms expand to uppercase.
   - The synonyms files cannot contain terms that are specified as stop words. For example, if `on` is included in your stop words file, and you specify in your synonyms file that `rotfl` expands to `rolling on the floor laughing`, the expansion won't return the expected results. Check the words in the stop words file that is used by your collection by default to make sure that you don't use any of the same words. For more information, see Identifying words to ignore.

   You can use the expansions.json file as a starting point when you build a query expansion list.

2. From the navigation pane, open the **Improve and customize** page.
3. Expand **Improve relevance** from the Improvement tools pane.
4. Click **Synonyms**, and then click **Upload synonyms** for the collection.

   Do not upload a synonyms file while documents are being added to your collection. The ingestion processing that occurs when documents are added can cause the index to be unavailable.

   Only one synonyms list can be uploaded per collection. If a second expansion list is uploaded, the second list replaces the first.

5. Run a test query to verify that the query expansion is working as expected.

   Query expansions are applied at query time, not during indexing, so you can add synonyms without reprocessing your collection.

To disable query expansion, delete the synonyms file. However, do not delete a synonyms file while new documents are being processed.

Identifying words to ignore

To ignore meaningless terms during searches, add a list of custom stop words. Stop words are words that are not useful in distinguishing the semantic meaning of the content.

In English, `the`, `is` and `and` are examples of stop words.

The stop words that you define are filtered out of queries and improve the relevance of natural language query results.

For example, a company has three tiers of service. The documents in one of the collections pertain to only one tier, the Silver tier. You might want to add `"silver"` to the stop words list because the term doesn't help to distinguish the significance of one document over another, given that all of the documents relate to the Silver service tier. When a customer mentions the Silver tier in a query string, it is ignored. Other terms in the query that are more significant are used to search the data instead. Or maybe the document collection consists of car accident reports only. You might want to add `"car"` to the stop words list to prevent mentions of `car` in queries from adding noise to the search.

Discovery applies a list of default stop words for many of the supported languages automatically. These stop words are applied both at indexing time and at query time. The predefined stop words are ignored when content is indexed and they are filtered out of queries. However, stop words that you define are used at query time only. Your list doesn't replace the default list; it augments the default list. You can add stop words, but you cannot remove stop words.

Example custom stop word list:

```
{
   "stopwords":  [
      "a",  "an",  "the",  "ibm",  "what",  "how",  "when",  "can",  "should",  ...
   ]
}
```

Default stop word lists

You can access the default stop words list for English from the Watson Developer Cloud GitHub repository.

For the following languages, Discovery uses the default stop words list that is defined by Apache Lucene. For more information about what words are included in the list, see the Lucene reference documentation:

- Arabic: stopwords_ar.txt
- Czech: stopwords_cs.txt
- Danish: stopwords_da.txt
- Dutch: stopwords_nl.txt
- Finnish: stopwords_fi.txt
- French: stopwords_fr.txt
- German: stopwords_de.txt
- Hindi: stopwords_hi.txt
- Italian: stopwords_it.txt
- Norwegian (both supported dialects): stopwords_no.txt
- Portuguese: stopwords_pt.txt
- Romanian: stopwords_ro.txt
- Russian: stopwords_ru.txt
- Spanish: stopwords_es.txt
- Swedish: stopwords_sv.txt
- Turkish: stopwords_tr.txt

Note: These default stop words are documented in TXT format, but if you want to augment the list and submit it for use by Discovery, you must submit a JSON file. To see an example of the syntax of stop words list file, see the custom English stop words list file.

For the remaining supported languages, no default stop words are used. You can specify a stop words list to use at query time for these languages. The list that you submit is not used when data is ingested.

Examples of stop word lists that you might want to apply at query time include:

- Japanese: custom_stopwords_ja.json
- Polish: custom_stopwords_pl.json

See supported languages for the list of the languages that are supported by Discovery.

Defining query-time stop words

To define stop words, complete the following steps:

1. Create a stop words file. The file must be a JSON file with the `json` file extension.

   Follow these guidelines:

   - Specify stop words in lowercase.

- In general, keep your list of stop words under `200` total words. The size limit is one million characters. However, if you specify too many terms, you might negatively affect search accuracy.

You can use the default English stop words list file, custom_stopwords_en.json, as a starting point when you build a custom stop word list in English.

2. From the navigation pane, open the **Improve and customize** page.
3. Expand **Improve relevance** from the Improvement tools pane.
4. Click **Stopwords**, and then click **Upload stopwords** for the collection.

   Only one stop words list can be uploaded per collection. The stop words list that you upload augments the default stop words list for your collection; it does not replace the default list.
5. Click **Done**.

To disable a custom stop words file and revert to using the default stop words, delete the custom stop words file.

Split documents to make query results more succinct

Split your documents so that the search function can find more concise information to return in query results.

For more information about the benefits of splitting documents, read the Using IBM Watson Discovery's New Document Segmentation Feature blog post on Medium.com.

Note: You can split only documents to which a user-trained Smart Document Understanding model is applied.

When you split a document, the original document is broken into segments. Each segment contains a more uniform set of information. By splitting the content in your documents into segmented groups, you can enrich and index your data at a more granular level.

To control how your documents are split, you specify a field, such as `subtitle` or `question`, to use as the page break marker. The page break options are populated with fields that are created when you apply a user-trained Smart Document Understanding (SDU) model to the documents. For more information, see Using Smart Document Understanding. You cannot split documents with fields that are generated by a pretrained Smart Document Understanding model.

As a document is reprocessed, it is evaluated from start to end. Whenever the page break marker field occurs, the original document is split and a new segment is created. The splitting continues at each marker field until the original document is broken into multiple segments.

Before you begin, decide which field to use as the page break marker.

- You can use any of the fields that are indexed by default. To see your choices, check the *Fields to index* list. Fields that have a *Type* value are stored in the index.
- The number of segments per document is limited to `1,000`. After segment number `999` is created, any remaining document content is stored within segment `1,000`.
- Metadata from PDF and Microsoft Word documents and any custom metadata is extracted and included in the index with each segment.

Be careful with documents that contain repeating sections, such as a catalog that has a description and specifications section for each product entry. If you split the document at too granular a level, the subsections, such as a section with specification details, can be disassociated from the product to which it belongs.

To split the documents in a collection, complete the following steps:

1. Click **Manage collections** from the navigation panel, and then click to open a collection.
2. Open the **Manage fields** page.

   A list of the identified fields is displayed.
3. From the *Improve query results by splitting your documents* section, click **Split document**.
4. Choose the field that you want to use as your page break marker from the **Select field** dropdown.

   The list that you can choose from includes a subset of all the identified fields.
5. Click **Apply changes and reprocess**.

You can check the status of the splitting process from the *Activity* page.

The metadata field includes the parent document ID. Each resulting segment of the original document can contain different information. For example, if you split the document based on the subtitle field, the first segment might contain only a title field. The next segment might contain a subtitle and a text field. The third might contain a subtitle field, a text field, and a footer field.

Updating documents that were split

If a document that was split changes and you want to upload the document again, work with a developer to replace the document by using the API. A developer can use the *Update a document* method to replace the original parent document. For more information, see the API reference. To provide the `{document_id}` path variable that must be sent with the request, copy the contents of the `parent_document_id` field of one of the document's segments.

When you replace the original document, all of the segments are overwritten, unless the updated version of the document has fewer total segments than the original. Those older segments remain in the index.

Deleting document segments from the index

You can delete documents in a collection from the *Manage data* page. To find all of the document segments that were generated from a single document, check for documents with the same `metadata.parent_document_id` field value. For more information, see Excluding content from query results.

IBM Cloud Pak for Data **IBM Cloud Pak for Data before the 4.6.5 release**

The *Manage data* page is available in installed deployments starting with the 4.6.5 release. In earlier releases, a developer can delete document segments by using the API. For more information, see the delete document API.

Excluding content from query results

Prevent content that you don't want customers to see from being included in query results.

You can prevent content from being included in query results in the following ways:

- Delete an entire collection.

  For more information, see Deleting collections.

- Remove a field from the index that contains data that you don't want to share with customers.

  You can control which fields are indexed. If you want to prevent a field from being indexed, you can set it to be excluded. For example, if your PDF files contain a running header or footer that does not contain useful information, you can exclude the `header` and `footer` fields from the index.

  To manage the fields to index, complete the following steps:

  1. From the navigation pane, open the **Manage collections** page, and then click a collection to open it.
  2. Click the **Manage fields** tab.

     A list of the identified fields is displayed. You can see which fields are included in the index and which are not.
  3. To remove a field from the index, set the **Include** switch to off.

- Delete a single document.

  > **Note:** If you use the Smart Document Understanding tool to annotate a document, and then decide that you want to delete the document and its associated SDU annotations, you must remove the annotations before you delete the document. To remove the annotations, annotate the document again. This time, label all of the content as `text`.

  To delete a document, complete the following steps:

  1. From the navigation pane, open the **Manage collections** page, and then click a collection to open it.
  2. Click the **Manage data** tab.

     A list of information from each document in the collection is displayed. If the information that is displayed doesn't help you identify the document that you want to delete, you can change what is displayed.

     - Click the *Settings* icon in the table header.
     - Choose fields from which you want to fetch data to display in the first and second columns. You can choose fields such as `extracted_metadata.filename` to show the document file name, or `document_id`, for example.

     > **Tip:** You can page through the documents in the collection by using the controls in the table footer.

  3. After you identify the document that you want to delete, select the checkbox that is associated with the document, and then click **Delete**. Confirm the deletion.

     Documents that are added to a collection from an external data source will be added back to the collection with the next scheduled crawl of the data source. The delete function removes the document from the index of the collection, not from the external data source.

     > **Note:** Some file types, such as CSV or JSON files, generate subdocuments when they are added to a collection. Splitting a document turns one document into multiple document segments. If you delete one of these generated documents, and then repeat the action that

created it, the deleted document is added back in to your collection.

IBM Cloud Pak for Data **IBM Cloud Pak for Data releases before 4.6.5**

The *Manage data* page is not available in installed deployments before the 4.6.5 release. You must use the Discovery API to delete a document. And you must know the document ID of the document that you want to delete. To get the document ID, use the List documents API method.

If the document is a subdocument of another document and you want to remove it, its parent, and any other subdocuments that are associated with the parent, delete the parent document. To get the document ID of the parent document, look for the `metadata.parent_document_id` field for the document. It is specified in the JSON representation of the document when it is returned as a response in the *Improve and customize* page of the product user interface.

Improving result relevance with training

The relevance of natural language query results can be improved in IBM Watson® Discovery with training.

A relevancy model determines the most relevant documents to return in search results. Without relevancy training, a standard mechanism is used to determine relevance based on common factors. When you train a relevancy model, you help Discovery to use features that are unique to your documents as it determines relevance.

The relevancy training model that is associated with a project is used at run time only when natural language queries are submitted. The model is not applied to Discovery Query Language (DQL) queries.

⚠ **Important:** You cannot apply relevancy training to *Content Mining* project types.

To train a relevancy model, you provide sample natural language queries, submit them to get results from your documents, and then rate those results. As you add more examples, the information you provide about result relevance for each query is used to learn about your project. The system uses your assessments to assign importance to different types of structural information within the documents. For example, it learns the importance of when a keyword from the search query appears in the title versus the header, body, or in the metadata of the document. It also learns from the importance of the distance between one matching keyword and another. After a successful relevancy training session, a ranker model is created. The model is used automatically by Discovery with the next natural language query. Discovery reorders the document results so that the most relevant results according to the relevancy training model are displayed first.

Training applies to an entire project. It cannot be skipped for one collection and applied to other collections in the same project. You do not enable use of the training model by specifying a query parameter. If present, the model is used for every natural language query that is submitted for the project. The model is used whether you limit the search to one collection or all of the collections. For this reason, it is important that your training data represents queries that are likely to be answered by all of the collections in your project. To stop a project from using the relevancy training model, you can delete the model by using the API.

Relevancy training does not run continuously. Training occurs only when you initiate it. At most one trained relevancy model is used at a time per project. If you retrain a model, the existing model is used until the new model is successfully trained, at which time the new model replaces the old model.

The set of documents that constitute the training data are used only during the training process. If a subsequent change is made to a document that was used to train the model, it does not change the trained model and does not trigger a new training session. Keep in mind that if many of the documents in your project change, it might be time to retrain the model to use the features from the updated documents.

Stop words and query expansions that you add to a collection do not affect the relevancy training model directly. However, they can change which documents are returned from a search, which affects the documents that are ranked by the relevancy model. The model ranks the top 100 documents that are returned for a query. Changes that you make to stop words or query expansions do not initiate a relevancy training update. If you add artifacts that drastically change the documents that are returned by search, consider retraining the model.

If documents that were used previously to train the model are removed from a collection, you must remove any references to them from the training data before you start to retrain the model. The model expects both the documents and queries from training data pairs to continue to exist. To remove these references, delete the training queries that returned the deleted documents. If the queries continue to be relevant, you can add them back to the training data and pair them with other documents.

For more information about the relevancy training API, see the API reference documentation.

When to use relevancy training

Relevancy training is optional. Test the quality of your search results. If the results of your queries meet your needs, no further training is necessary.

The training improves the relevancy of the documents that are returned in query responses. It does not improve the passages or answers that are returned per document. If you're using passage retrieval and your test results are returning good documents, but the wrong passages from the documents, relevancy training will not help.

For more information about when to use relevancy training, read the Relevancy training for time-sensitive users blog post on Medium.

How fields are handled

When you train a project from the product user interface, the results are always taken from the `text` field of the documents. If your documents don't have a `text` field, use the API to train your project instead. Your documents might not have a `text` field if you uploaded a CSV file that doesn't have a column named *text*, or uploaded a JSON file that doesn't have an object named *text*, or if you used the Smart Document Understanding tool to define fields with other names in which the bulk of the content of your documents now are stored.

When you train a project from the API, results are taken from all of the root-level fields and they are all considered to have equal significance. Unlike Discovery Query Language queries, with natural language queries you cannot specify which fields from the document you care about or how much significance to give to each one. When you teach Discovery with examples, the service figures out for you how much weight to give to each field.

Discovery builds a model that assigns different weights to term, bigram, and skip-gram matches for each of the root-level fields and balances them against matches from all of the other document fields. With enough examples, Discovery can return better answers because it knows where the best answers are typically stored.

Note: Relevancy training cannot be used to give more weight to nested fields. Nested fields are grouped and assigned one overall score. No matter how much you train, Discovery never gives a nested field more weight than it gives to a root-level field. For more information about nested fields, see the FAQ.

Training a project

The training data that is used to train the relevancy model includes these parts:

- A natural language query that is representative of a query that your users might submit
- Results of the query which are returned by the service
- The rating that you apply to the result that indicates whether the result is `relevant` or `not relevant`

To apply relevancy training to a project, complete the following steps:

1. Go to the **Improve and customize** page. On the **Improvement tools** panel, select **Improve relevance**, then **Relevancy training**.
2. Enter a natural language query in the **Enter a question to train** field.

   Do not include a question mark in your query. Use the same wording as your users. For example, `IBM Watson in healthcare`. Write queries that include some of the terms that are mentioned in the target answer. Term overlap improves the initial results when the natural language query is evaluated.
3. Click **Add+**.
4. Click **Rate results**.
5. After the results are displayed, assess each result, and then select **Relevant** or **Not relevant**, whichever option applies given the quality of the result.

   When you select **Relevant**, you apply a score of `10` to the result. **Not relevant** applies a score of `0`. You can use a different scoring scale if you use the API to rate results, but you can't mix scoring scales within the same project.

   If the result shows the message, “No content preview available for this document”, it means that the document that was returned does not contain a `text` field or that its `text` field is empty. If none of the documents in your collection have a `text` field, use the API to train the project instead of training it from the product user interface.
6. When you are finished, click **Back to queries**.
7. Continue adding queries and rating them.

   As you rate results, your progress is shown. Check your progress to see when enough rating information is available to meet the training threshold needs. Your progress is broken into the following tasks:

   - Add more queries
   - Rate more results
   - Add more variety to your ratings

   You must evaluate at least 50 unique queries, maybe more, depending on the complexity of your data. You cannot add more than 10,000 training queries.
8. You can continue adding queries and rating results after you reach the threshold. Enter all of the queries that you think your users will ask.

   To delete a training query, click the **Delete** icon. To delete all of the training queries in your collection at one time, use the API. For more information, see Delete training queries.

Note: If two or more users attempt to train identical queries at the same time, the ratings that are submitted by one of the users overwrites the others.

Testing and iterating on the relevancy of results

When you are done rating results, and training is completed, test to see whether your query results are better. To do so, run test natural language queries that are related (but not identical) to your training queries. Review the results.

If you want to continue to improve the results after testing, you can:

- Add more documents to your collection.
- Add more training queries.
- Rate more results, making sure to use both the `Relevant` and `Not relevant` ratings.

Confidence scores

Discovery returns a `confidence` score for natural language queries of trained collections. This `confidence` score is not interchangeable with `confidence` scores that are returned by untrained collections.

The `confidence` score can range from `0.0` to `1.0`. The higher the number, the more relevant the result.

The `confidence` score can be found in the query results, under the `result_metadata` for each document. This number is calculated based on how relevant the result is estimated to be, compared to the trained model.

```
{
   "matching_results":  4,
   "retrieval_details":  {
      "document_retrieval_strategy":  "trained"
   },
   "results":  [
      {
    "id":  "eea16dfd5fe6139a25324e7481a32f89_13",
    "result_metadata":  {
       "confidence":  0.08793
    }
      }
   ]
}
```

The `document_retrieval_strategy` can be found under the `retrieval_details`. If you query a trained collection by using the Discovery Query Language, or the trained model is temporarily disabled, the `document_retrieval_strategy` is `untrained`.

For more information on querying a project, see the Query overview.

Relevancy training limits

The following limits apply to relevancy training models:

- One model per project
- 10,000 queries per model
- 40 models per service instance for Enterprise and Premium plans; 20 models for Plus plan instances

Other ways to improve relevancy

If you prefer to use the Discovery API to train Discovery, see the API reference.

You also can use the API to add curations. Curations is a beta feature that you can use to teach Discovery to return a specific document every time a certain query is submitted. For more information, see Curations.

Adding a custom stopwords list can also improve the relevance of results for natural language queries. For more information, see Identifying words to ignore.

Understanding relevancy training

Answers to common questions about training a project.

How do I know whether my system is trained?

Run a natural language query and check the `document_retrieval_strategy`. See confidence scores.

If you are using the API, see List training queries.

How long does it take to train a model?

It can take between 45 minutes to an hour for the training to finish. The duration of the training differs depending on the amount and variety of the data that is used to train the relevancy model. Also, the training occurs asynchronously. It can be delayed if other data that it needs is unavailable because it is being searched or processed in some other way.

How do I stop relevancy training from being applied to my project?

Use the API to delete the relevancy model that is associated with your project. To delete the model, you delete that training data that is associated with the ranker model. For more information, see Deleting training queries.

Does relevancy training impact passage search?

No. Relevancy training is used for document search only. It has no impact on passage search.

Does relevancy training impact answer finding?

Not directly. Relevancy training indirectly impacts answer finding because it changes the order of the documents from which answers are retrieved. It reranks the returned documents from most to least relevant.

How do I check errors and warnings?

Open the **Manage collections** page. Choose your collection, then open the **Activity** tab.

How do I interpret the confidence score that appears in natural language query results after training?

See confidence scores.

Interpreting relevancy training errors and warnings

The following list has explanations for some common error and warning messages.

Warning: `Invalid training data found: The document was not returned in the top 100 search results for the given query, and will not be used for training`

This warning occurs when the `document_ids` in your training data do not match the `document_ids` in a search that is performed against the collection. Check your queries to make sure that the `document_id` of the document you are rating is returned in the top 100 results for that query. If it is not, then you might want to check two things:

- If the document is not returned in the top 100, it might not be an example of a high-quality result. Reevaluate whether to use the document.
- If the document is not returned at all, then review why it is not returned and see whether any text in the document matches portions of the query.

> **Note:** This warning indicates that you might have one or more failed queries. It doesn't mean that the training cannot be completed.

Error: `Invalid training data found: Syntax error when parsing query`

A syntax error means that the query is invalid. Syntax errors can occur when you increase the complexity of the query by adding a filter to the natural language query, for example. Run the query against the collection outside of relevancy training by using the API. After you confirm that the query is valid and returns results, you can add it as a relevancy training query.

Error: `Training data quality standards not met: You will need additional training queries with labeled examples. (To be considered for training, each example must appear in the top 100 search results for its query.)`

You need to add more training data to train successfully. You need at least 49 unique training queries at a minimum, and each one needs at least one rated document. Minimum does not mean optimal; the size of the collection and other factors can increase the number of training examples that are needed to meet the minimum.

Error: `Training data quality standards not met: Insufficient number of unique training queries. Expected at least n, but found m.`

To meet the minimum training requirements, you need at least 50 unique training queries, and each query must have at least one rated document. If you have more queries than the minimum and are still receiving this error message, check your notices for other errors.

Error: `Training data quality standards not met: No documents found with non-zero relevance labels.`

Training data needs enough labeled data that specifies what documents are high value. Therefore, you need to rate some documents with nonzero values. You need to rate some documents as `Relevant` and some as `Not relevant`. At least one document must be rated `Relevant`.

Error: `Training data quality standards not met: Training examples have no relevance label variety for X queries.`

One of the requirements for training is to have sufficient label diversity. At least 25% of the training queries must include both `Relevant` and `Not relevant` labels. If you use the API, at least 25% of the queries must include two different numeric labels.

Default query settings

Learn about how the search query is configured for each project type by default.

When you submit a search from the product user interface, your text is passed as a natural language query value to the Query API. Other parameters that you can define when you use the API are assigned default values for queries that are made from the user interface. The following tables explain which values are specified by default for each project type. For more information about the Query API, see Query reference.

You can override some of the default values by using improvement tools in the user interface. For example, you can use the *Search results* tool to change parameters such as `passages.enabled`. For more information, see Changing the result content.

The enrichments that are applied to your data autmoatically differ by project type. For more information, see Default project settings.

Default query settings

Query default	Document Retrieval	Document Retrieval for Contracts
`aggregation`	`[term(enriched_text.entities.text,name:entities)]`	`[term(enriched_html.contract.elements.categories.label,count:25`
`count`	`10`	`10`
`highlight`	`false`	`false`
`passages.characters`	`200`	`200`
`passages.count`	`10`	`10`
`passages.enabled`	`true`	`true`
`passages.fields`	`["text", "title"]`	`["text", "title"]`
`passages.find_answers`	`false`	`false`
`passages.max_answers_per_passage`	`1`	`1`
`passages.max_per_document`	`1`	`1`
`passages.per_document`	`true`	`true`
`return`	`[]`	`[]`
`spelling_suggestions`	`false`	`true`
`sort`	`""`	`""`
`table_results.count`	`10`	`10`
`table_results.enabled`	`false`	`true`
`table_results.per_document`	`0`	`0`

Default query settings

Default query settings continued

Query default	Conversational Search	Content Mining	Custom
`aggregation`	`[ ]`	`[ ]`	`[ ]`
`count`	`10`	`10`	`10`
`highlight`	`false`	`false`	`false`
`passages.characters`	`200`	`200`	`200`

passages.count	10	10	10
passages.enabled	true	false	true
passages.fields	["text", "title"]	["text", "title"]	["text", "title"]
passages.find_answers	false	false	false
passages.max_answers_per_passage	1	1	1
passages.max_per_document	1	1	1
passages.per_document	true	true	true
return	[]	[]	[]
spelling_suggestions	false	true	true
sort	""	""	""
table_results.count	10	10	10
table_results.enabled	false	false	false
table_results.per_document	0	0	0

Default query settings continued

Project component settings

Default	Document Retrieval	Document Retrieval for Contracts
Aggregations	See table	See table
autocomplete	true	true
fields_shown.body.field	""	""
fields_shown.body.use_passage	true	true
fields_shown.title.field	"title"	"title"
results_per_page	5	5
structured_search	false	false

Project component settings

Project component settings continued

Note: The Custom project type has no project component default settings.

Default	Conversational search	Content Mining
Aggregations	[]	[]
autocomplete	false	true
fields_shown.body.field	""	text

`fields_shown.body.use_passage`	`true`	`false`
`fields_shown.title.field`	`"title"`	`"document_id"`
`results_per_page`	`0`	`0`
`structured_search`	`false`	`false`

Project component settings continued

Document Retrieval project aggregations

aggregations.name	aggregations.label	aggregations.multiple_selections_allowed
`"name": "entities"`	`"label": "Top Entities"`	`"multiple_selections_allowed": false`
`"name": "_system_collections"`	`"label": "Collections"`	`"multiple_selections_allowed": true`

Document Retrieval project aggregations

Document Retrieval for Contracts project aggregations

aggregations.name	aggregations.label	aggregations.multiple_selections_allowed
`"name": "categories"`	`"label": "Category"`	`"multiple_selections_allowed": true`
`"name": "natures"`	`"label": "Nature"`	`"multiple_selections_allowed": false`
`"name": "contract_terms"`	`"label": "Contract Term"`	`"multiple_selections_allowed": false`
`"name": "contract_payment_terms"`	`"label": "Contract Payment Term"`	`"multiple_selections_allowed": false`
`"name": "contract_types"`	`"label": "Contract Type"`	`"multiple_selections_allowed": false`
`"name": "contract_currencies"`	`"label": "Contract Currency"`	`"multiple_selections_allowed": false`
`"name": "invoice_buyers"`	`"label": "Invoice Buyer"`	`"multiple_selections_allowed": false`
`"name": "invoice_suppliers"`	`"label": "Invoice Supplier"`	`"multiple_selections_allowed": false`
`"name": "invoice_currencies"`	`"label": "Invoice Currency"`	`"multiple_selections_allowed": false`
`"name": "po_payment_terms"`	`"label": "Purchase Order Payment Term"`	`"multiple_selections_allowed": false`
`"name": "po_buyers"`	`"label": "Purchase Order Buyer"`	`"multiple_selections_allowed": false`
`"name": "po_suppliers"`	`"label": "Purchase Order Supplier"`	`"multiple_selections_allowed": false`
`"name": "po_currencies"`	`"label": "Purchase Order Currency"`	`"multiple_selections_allowed": false`

Document Retrieval for Contracts project aggregations

Enriching your data

Choose enrichments

Add resources that can teach Discovery about terms or patterns that have special meaning to your application.

The following table describes the best resources to add to address different needs.

Goal	Resource	Notes
Define categories by which text in your documents can be classified.	Classifier	N/A
Recognize terms and synonyms for terms that are significant to you, such as the names of products that you sell.	Dictionary	N/A
Define regular expressions that capture patterns of significance, such as that `AB10045` is the syntax that is used for your order numbers.	Regular expressions	N/A
Recognize and tag entities and relationships that are defined in a custom machine learning model.	Machine learning models	Requires a model that is built and exported from another IBM tool.
Apply rules to fields that are based on rules you defined by creating an advanced rules model in IBM Watson® Knowledge Studio.	Advanced rules models	Requires an advanced rules model that is built and exported from IBM Watson® Knowledge Studio or that uses an exported Patterns resource.
IBM Cloud Recognize terms that are mentioned in sentences that match a syntactic pattern that you teach Discovery to recognize.	Patterns (beta)	Available as a beta feature for English-language collections in managed deployments only. The enrichment that is derived by defining patterns cannot be applied to Content Mining projects. You can export the resource and use it as an advanced rules model.
Recognizes entities that you identify as being significant by training an entity extractor machine learning model.	Entity extractor	Supports starting from an imported Knowledge Studio corpus.
Classify sentences in your documents into user-defined sentence classes.	Sentence classifier	Supports smart labeling to speed up the labeling process.

Domain tools overview

Alternatively, you can apply built-in Watson NLP enrichments that find the following information in your collection:

- Entities and keywords
- Sentiment

You can extract meaning from documents based on the document structure by defining a Smart Document Understanding (SDU) model. Use the Smart Document Understanding tool to identify new fields by which to target enrichments or to split large documents into more manageable chunks. For more information, see Structural meaning with SDU.

Dictionaries and classifiers that you add to one project can be used by other projects.

For more information about how to get the most from enrichments, read the Enriching your documents can make search more effective blog post.

Choosing the right enrichment type

The following diagram helps you to choose the right enrichment for your use case.

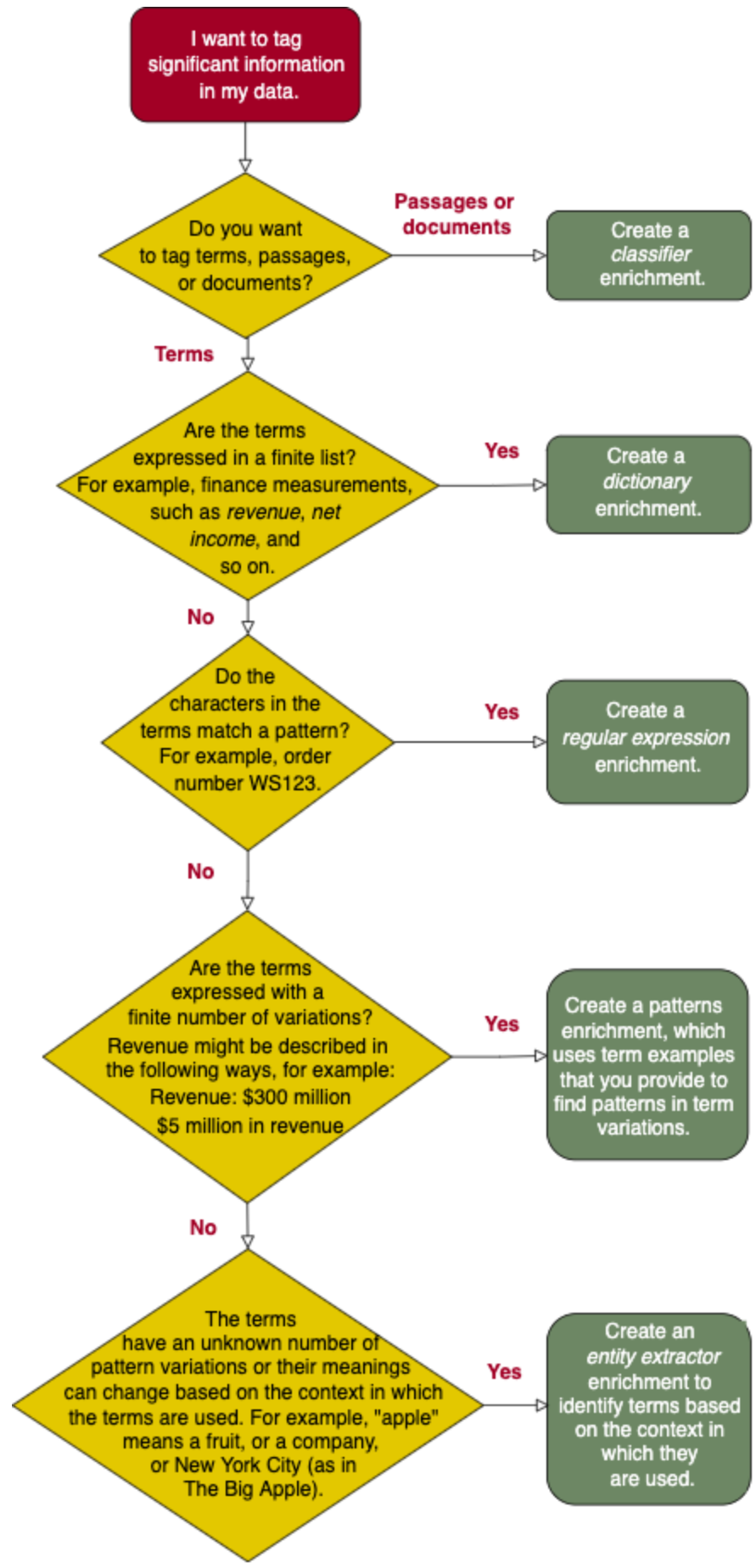

Figure 1. Flow diagram for choosing the right enrichment

Using enrichments together

You can use many enrichments together to tackle various challenges that you might encounter as you develop a search application.

Many teams start by creating a **dictionary** enrichment. Dictionaries are a great tool for identifying important terms and tagging them so they can be retrieved later. Let's say you're building a search application that needs to extract ingredients from recipes. A dictionary enrichment can recognize mentions of most ingredients. However, the dictionary enrichment might partially match against two-word terms. For terms such as `olive oil` or `mustard greens`, it might incorrectly recognize only `olive` and `mustard`. To improve the accuracy of the search, you can augment the dictionary enrichment with a **pattern** enrichment that can recognize two-word ingredient mentions. Maybe a few recipes mention food coloring codes in European format (`E104`). You can add a **regular expression** enrichment to recognize occurrences of codes with the syntax `E1nn`. Finally, to catch terms that no other enrichment can recognize, you can use a **machine learning** enrichment. The enrichment can be one that you build in an external tool and import to Discovery or one that you build in Discovery by creating an **entity extractor** enrichment.

The entity extractor enrichment is more sophisticated than the other enrichments. For example, a dictionary enrichment recognizes only exact matches of dictionary terms and synonyms that occur in your documents. A regular expression enrichment recognizes only specific patterns. In contrast, occurrences of an entity are recognized based on the context in which an entity example is mentioned in a sentence.

For example, maybe you want to recognize locations and the document you want to process contains the following types of sentences:

- I live in `Massachusetts`.
- We're traveling from `New York City` to `Paris` next week.

To use a dictionary enrichment to recognize location names successfully, the dictionary must list every possible location. However, if you use an entity extractor enrichment, you can identify when a location is mentioned based on how the location is referenced in a sentence. With phrases such as, “I live in `x`” or “I'm from `x`” or “I'm traveling to `x`” in its training data, the entity extractor can learn that `x` is a reference to a location.

When you need to choose between using a dictionary or an entity extractor enrichment, follow these guidelines:

- If the list of possible examples is short, use a dictionary.

  It is more efficient to define a dictionary term `planet` with synonyms such as `Earth` and `Saturn` than to create a `planet` entity because only 8 planets exist in our solar system. However, defining a list of every possible location on Earth is not feasible. An entity extractor can recognize more location mentions.

- If the list of possible examples is static, use a dictionary.

  Controversy over Pluto aside, the `planet` category is a good example here too because the list of planets in our solar system is static. Or maybe you want to monitor general customer sentiment about your products. You need to be able to recognize product name mentions, but might not need specifics. If you have a large variety of product names, you can create a `product name` entity. As new products are added to your portfolio, or product names change over time, you do not need to maintain an overall product list. The entity extractor can continue to recognize general feedback about your products based on the context of the sentences in which products are mentioned.

Add a resource

When you add a custom enrichment to a project it is available to any collection in the project.

To add a resource, complete the following steps:

1. Open your project and go to the *Improve and customize* page.
2. On the **Improvement tools** panel, expand **Teach domain concepts**, and then choose the resource that you want to add.

   After you create the resource, it becomes a new type of enrichment that you can apply to your data.

3. Specify the collection and field in which to apply the enrichment.

   > **Tip:** You can apply enrichments to the `text` and `html` fields, and to custom fields that were added from uploaded JSON or CSV files or from the Smart Document Understanding (SDU) tool. Only the first 50,000 characters of a custom field from a JSON file are enriched.

   For example, if you add a dictionary and choose to apply it to the `text` field of a collection, the documents in the collection are reprocessed. If the term `vehicle` is specified as a synonym of the `car` dictionary entry and occurs in the document text, `vehicle` is tagged as a mention of the `car` dictionary entry type. If a customer later searches for `car`, the passage that contains the `vehicle` mention is included in the search results.

   > **Note:** If the field that you choose comes from a JSON file, after you apply the enrichment, the field data type is converted to an array. The field is converted to an array even if it contains a single value. For example, `"field1": "Discovery"` becomes `"field1": ["Discovery"]`.

You can choose to apply resource-derived enrichments to your data later. Enrichments that you add to a project are available for use from any collection in the project. Go to the *Manage collections* page, choose the collection where you want to apply the enrichment, and then open the **Enrichments** tab. Make sure the status of the enrichment shows that it is *Ready*, and then apply the enrichment to a field in the collection. Enrichments that you enable are applied to the documents in random order. For more information, see Managing enrichments.

From the deployed Content Mining application, you can create a classifier or a custom annotator from a dictionary, regular expression, machine learning, or PEAR file and use it as an enrichment in collections that are stored in other project types. For more information, see Adding facets.

Use built-in Watson NLP to find common terms

Take advantage of award-winning Watson Natural Language Processing (NLP) capabilities by adding prebuilt enrichments to your documents.

With Watson NLP, you can identify and tag meaningful information in your collections so you can understand what it all means and make more informed decisions.

The following Watson NLP enrichments are available:

- Entities: Recognizes proper nouns such as people, cities, and organizations that are mentioned in the content.
- Keywords: Recognizes significant terms in your content.
- Part of Speech: Identifies the parts of speech (nouns and verbs, for example) in the content.
- Sentiment: Understands the overall sentiment of the content.

The following other pretrained enrichments are available with Discovery:

- Contracts
- Document structure
- Table understanding

Watson NLP enrichments

For example, the following screen capture shows a transcript of the US Declaration of Independence that was added to a Discovery collection where the Entities and Keywords enrichments are enabled. The mentions that are recognized by the enrichments are highlighted in the document text.

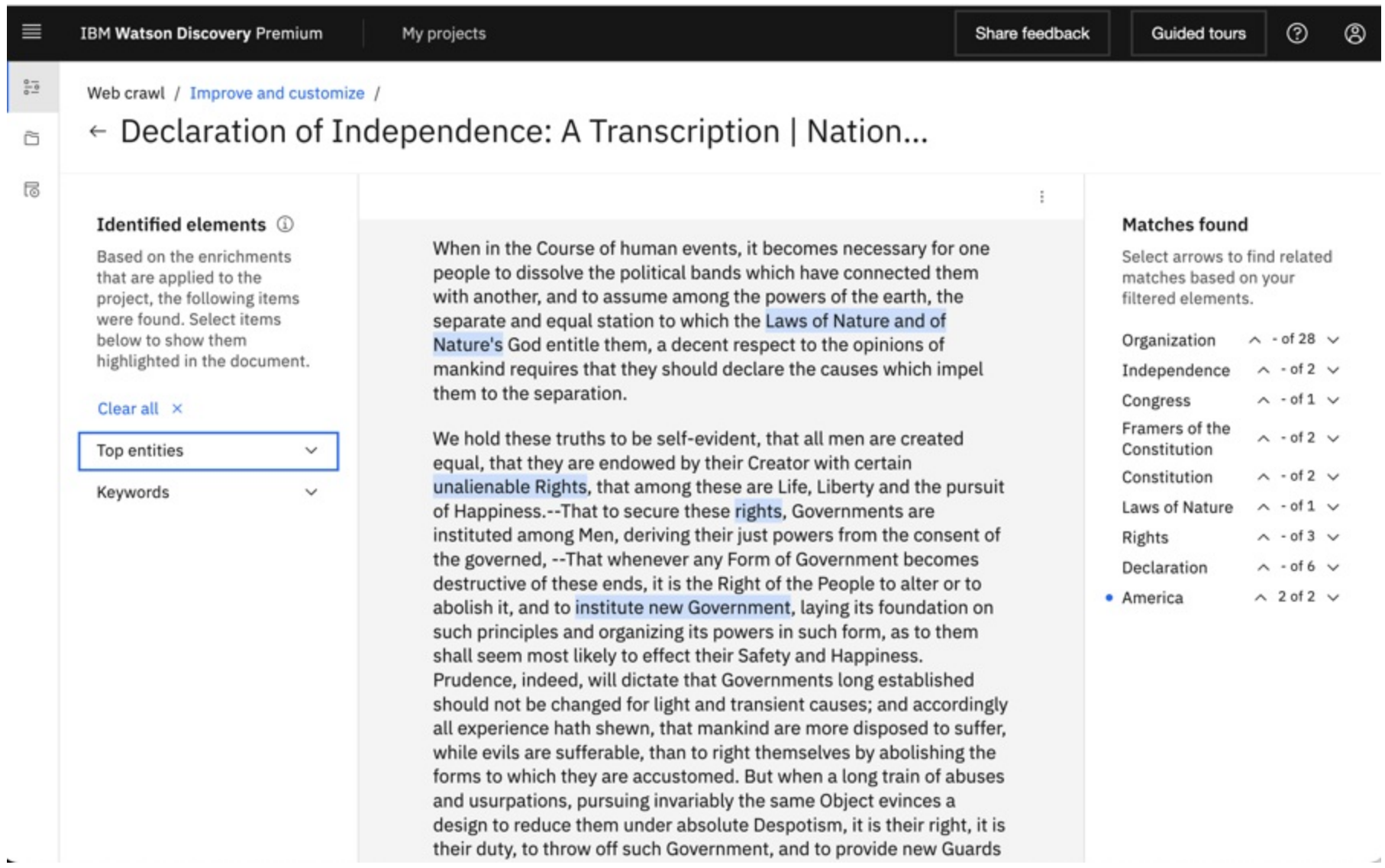

Figure 1. Excerpt of the US Declaration of Independence with highlighted terms

Some of the NLP enrichments are applied to projects automatically. You don't need to apply them yourself if you are using one of these project types.

Default enrichments per project type

Some prebuilt enrichments are applied automatically to collections in a project based on the project type. The following table shows the default enrichments that are applied to each project type.

Enrichment	Document Retrieval	Document Retrieval for Contracts	Conversational Search	Content Mining
Contracts		✓		
Entities	✓	✓		
Keywords				
Part of Speech			✓	
Sentiment of Document				
Table Understanding		✓		

Default enrichments per project type

For more information about the following prebuilt enrichments, see the following topics:

- Contracts
- Table Understanding

For more information about how to create custom enrichments, see Adding domain-specific resources.

For more information about how to get the most from enrichments, read the Enriching your documents can make search more effective blog post.

For more information about how to apply enrichments by using the API, see Applying enrichments by using the API.

Add enrichments

To add an NLP enrichment, complete the following steps:

1. Open your project and go to the *Manage collections* page.
2. Click to open the collection that you want to enrich.
3. Open the **Enrichments** tab.
4. Scroll to find the NLP enrichment that you want to apply to your documents.

   > **Note:** Both built-in enrichments and custom enrichments are listed. Built-in enrichments have a type value of `System`.

5. Choose one or more fields to apply the enrichment to.

   You can apply enrichments to the `text` and `html` fields, and to custom fields that were added from uploaded JSON or CSV files or from the Smart Document Understanding (SDU) tool.

6. Click **Apply changes and reprocess**.

Enrichments that you enable are applied to the documents in random order. For information about how to remove an enrichment, see Managing enrichments.

Entities

Identifies entities. *Entities* are terms that typically represent proper nouns such as people, cities, and organizations that are mentioned in the data collection. Discovery can recognize entities that are part of an entity type system that is defined by the Watson Natural Language Processing (NLP) service.

If you want to be able to identify uncommon terms that are significant to your business, you can train your own model to recognize custom entities. For more information, see Entity extractor.

The Watson NLP entity extractor service that is used by Discovery is called the *NLU type system*. The name originates from the fact that the type system is used by the Watson Natural Language Understanding (NLU) service in addition to the Watson Discovery service. However, it is the Watson NLP implementation of the type system that is used directly by Discovery, not the Watson NLU implementation. As a result, the two implementations can produce different results. To get a general idea of the types of entities that are recognized by the service, see Entities.

The following screen capture shows that the Entities enrichment recognizes the terms *Systems of Government* and *King of Great Britain* (among others) and tags them as entity mentions.

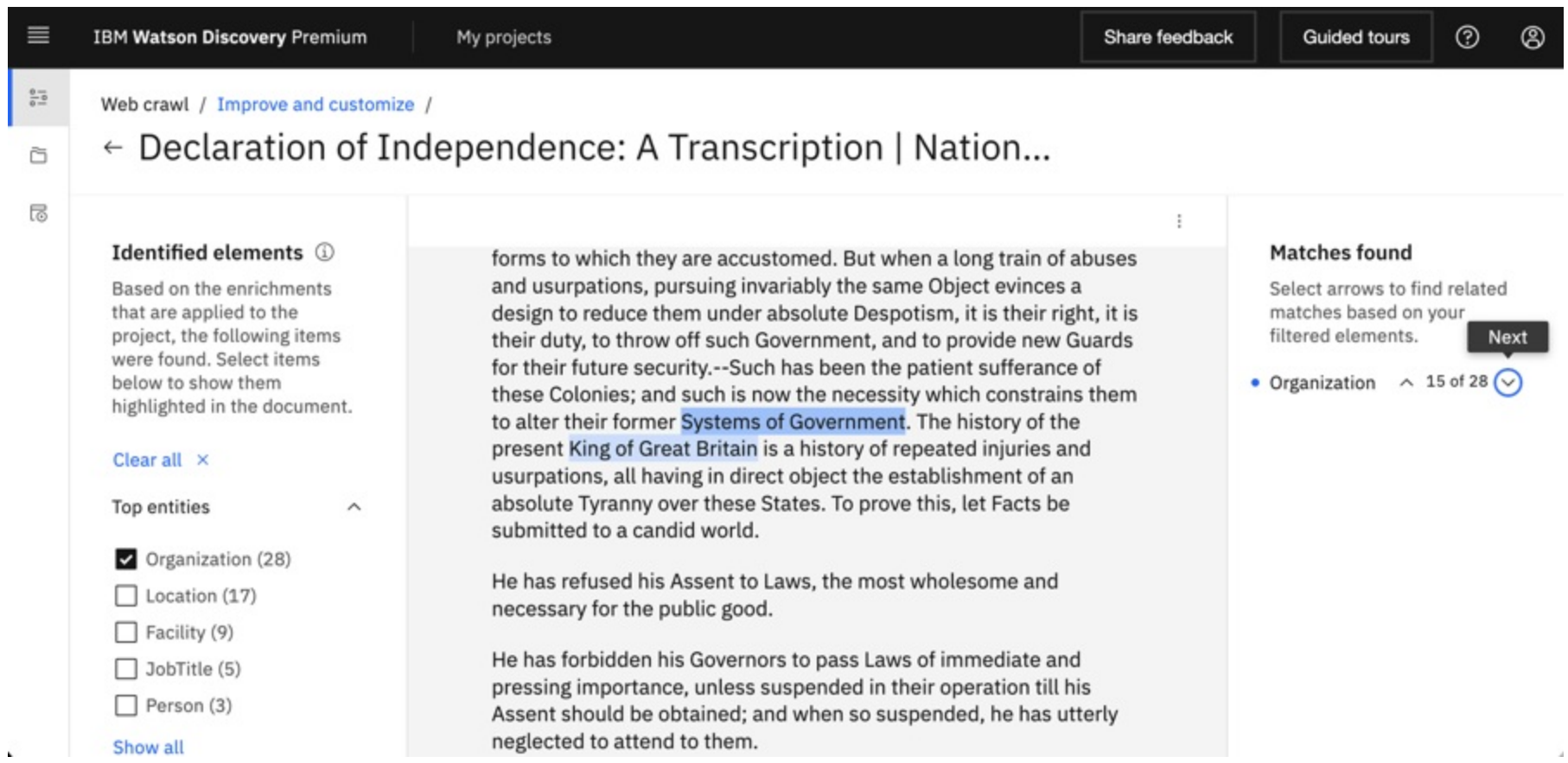

Figure 2. The recognized entities, Governments and King of Great Britain, are highlighted

From the JSON view of the document, you can see the underlying JSON structure of the entity mentions.

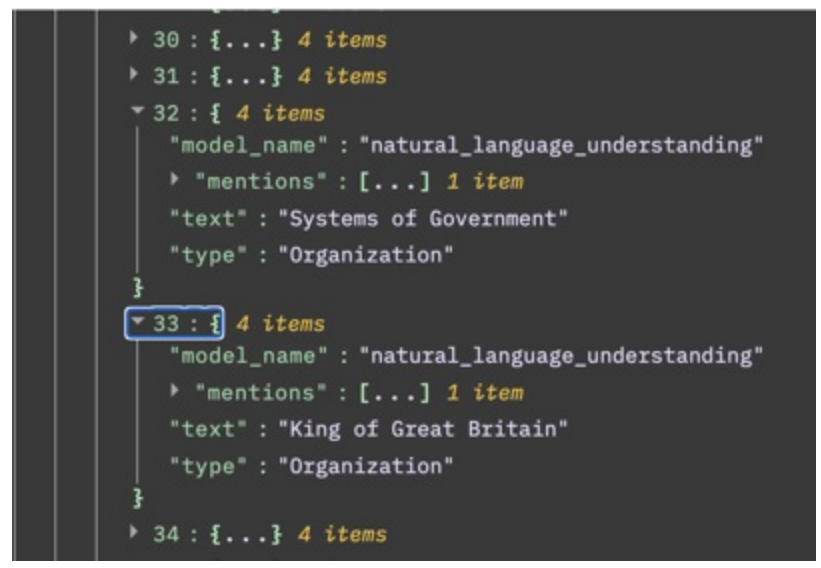

Figure 3. JSON representation of recognized entity mentions

If you want to search for the Organization entity type, for example, you can copy all of the JSON content into a text editor and search for `Organization`. Click the *Copy* icon from the root of the JSON tree view.

Example

Input

```
"IBM is an American multinational technology company headquartered in Armonk."
```

Response

In the JSON output:

- `text` = string. The entity text
- `type` = string. The entity type, such as `Organization`, `Location`, `Person`, `Number`.
- `mentions` = array. The entity mentions and locations
- `model_name` = string. For custom models, this field contains the user-provided model name. Otherwise, this field contains the default name of the model, such as `watson_knowledge_studio`, `dictionary`, `character_pattern`, or `natural_language_understanding`


```
{
   "entities":  [
      {
         "model_name":  "natural_language_understanding",
         "mentions":  [
            {
               "confidence":  0.8317045,
               "location":  {
                  "end":  3,
                  "begin":  0
               },
               "text":  "IBM"
            }
         ],
         "text":  "IBM",
         "type":  "Organization"
      },
      {
         "model_name":  "natural_language_understanding",
         "mentions":  [
            {
               "confidence":  0.6114863,
               "location":  {
                  "end":  75,
                  "begin":  69
               },
            "text":  "Armonk"
            }
         ],
         "text":  "Armonk",
         "type":  "Location"
      }
   ]
}
```


Entity limits

The Entities enrichment can identify up to 50 entities, each with one or many mentions, per document.

Keywords

Returns important keywords in the content.

For example, the following screen capture shows highlighted terms from the US Declaration of Independence that are recognized by the Keywords enrichment.

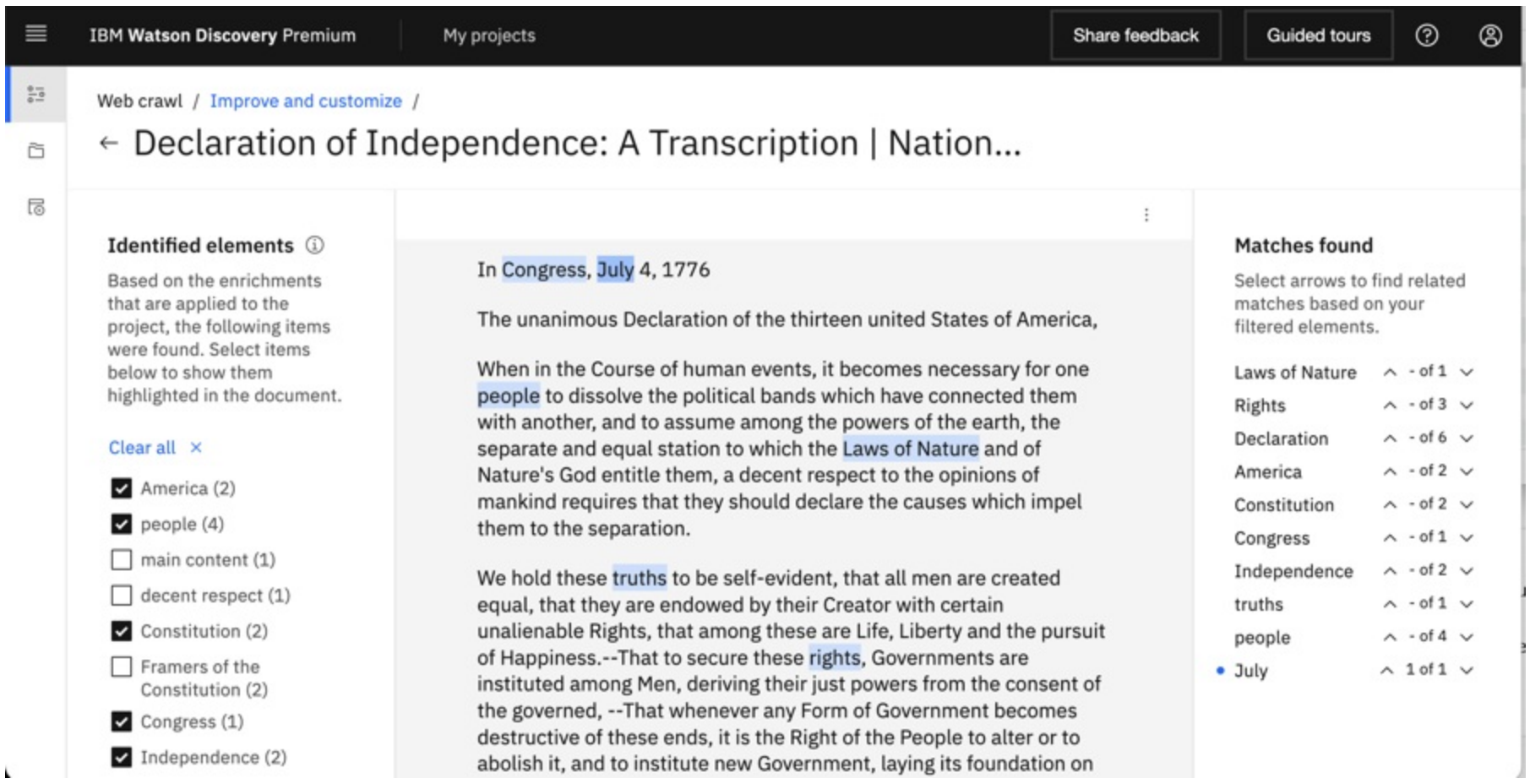

Figure 4. Terms recognized by the Keywords enrichment

From the JSON view of the document, you can see the underlying JSON structure of the `Declaration` keyword mention.

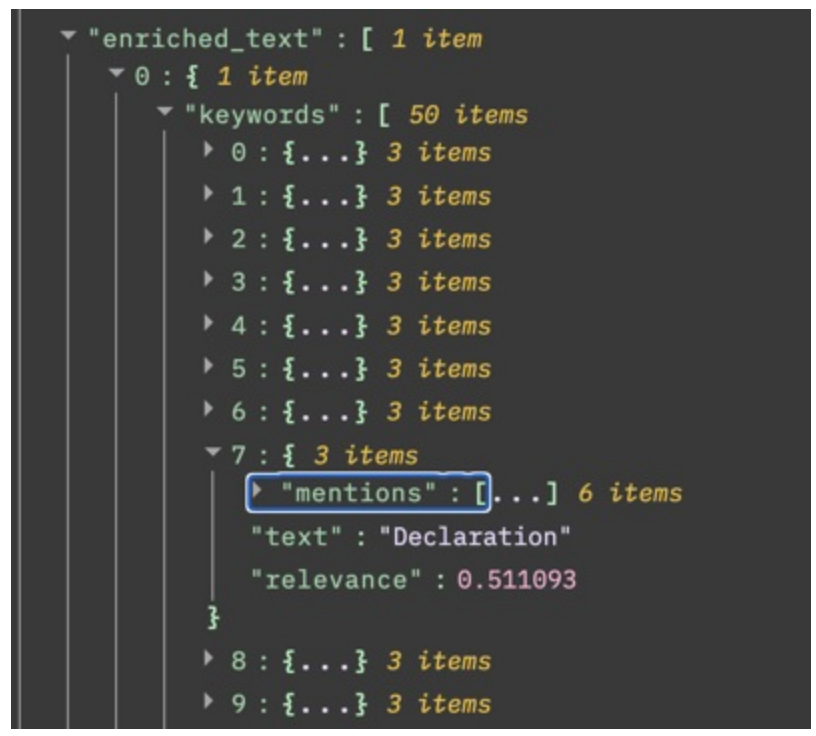

Figure 5. JSON representation of Keywords enrichment mentions

Example

Input

```
"Watson Discovery is an award-winning AI search technology."
```

Response

In the JSON output:

- `text` = The keyword text
- `mentions` = The entity mentions and locations

```
{
   "keywords":  [
      {
         "mentions":  [
            {
               "location":  {
```

```
                "end":  157,
                "begin":  141
             },
             "text":  "Watson Discovery"
          }
       ],
       "text":  "Watson Discovery",
       "relevance":  0.503613
    },
    {
       "mentions":  [
          {
             "location":  {
               "end":  177,
                "begin":  164
             },
             "text":  "award-winning"
          }
       ],
       "text":  "award-winning",
       "relevance":  0.728722
    },
    {
       "mentions":  [
          {
             "location":  {
                "end":  198,
                "begin":  181
             },
             "text":  "search technology"
          }
       ],
       "text":  "search technology",
       "relevance":  0.779356
    }
  ]
}
```


Keywords limits

The Keywords enrichment can identify up to 50 keywords, each with one or many mentions, per document.

Part of speech

Recognizes and tags parts of speech, including nouns, verbs, adjectives, adverbs, conjunctions, interjections, and numerals.

Identify custom terms

Define a finite set of terms with a dictionary

Recognize terms and synonyms for terms that are significant to you, such as the names of products that you sell.

Help Discovery find terms that have meaning to your use case by adding a dictionary. You can define multiple synonyms for a term or a set of words in the same category.

You can create a dictionary by adding the terms one by one or by uploading a CSV file that lists the terms.

To add dictionary terms one by one, complete the following steps:

1. From the *Teach domain concepts* section of the *Improvement tools* panel, choose **Dictionaries**.
2. Click **New**.
3. Name your dictionary.

   For example, `Transportation`.
4. Choose the language. A dictionary can contain terms in only one language.
5. **Optional**: Expand *Advanced options*, and edit the facet name for the dictionary.

   Facets are used to categorize documents. A user can choose a facet type to narrow their search results. The dictionary name in lowercase is used as the facet name by default. You might want to change the facet to be uppercase.
6. Enter a term, and then select the **+** button to add it.

For example `vehicle` and `engine`.

In English dictionaries, specify the dictionary terms in lowercase. Only use uppercase if you want Discovery to ignore lowercase mentions of the term when they occur in text. When terms are analyzed to determine whether they are occurrences of the dictionary enrichment, the surface form of the term with uppercase match is used. For example, a `vehicle` entry in the dictionary results in annotations for `vehicle`, `Vehicle`, or `VEHICLE` mentions when they occur in text. For a `Sat` entry in the dictionary, annotations are added for `Sat` or `SAT`, but not for `sat`.

Dictionary matching is case-sensitive for Arabic, Chinese, Korean, Japanese, and Hebrew.

7. To add synonyms for the term, click the *Edit* icon, and then enter synonyms in the **Other terms** field. Separate multiple synonyms with a comma. Click **Save term**.

   The dictionary can contain terms and their synonyms or a category and terms that belong to the category.

   For the term `vehicle`, you can specify synonyms such as `car`, `automobile`, `sedan`, `convertible`, `station wagon`, and so on. For `engine`, you can specify `gasket`, `carburetor`, `piston`, and `valves`.

   > **Tip:** Be careful not to add too many synonyms. Test the impact of any synonyms that you add. When you test, use data that is different from the data you use to derive the synonyms.

8. Continue adding terms.

   Similar terms from all of the collections in the current project are suggested as new entries.

   > **Note:** Suggested terms are taken from a field named *text*. If the text field is missing, a field with the longest string value and highest number of distinct values is chosen. Suggestions are not displayed if there are no documents or the collection has no fields with text data.

9. Click **Save dictionary**.

10. Choose the collections and fields where you want to apply the dictionary, and then click **Apply**.

Example

A transportation dictionary is added to a project.

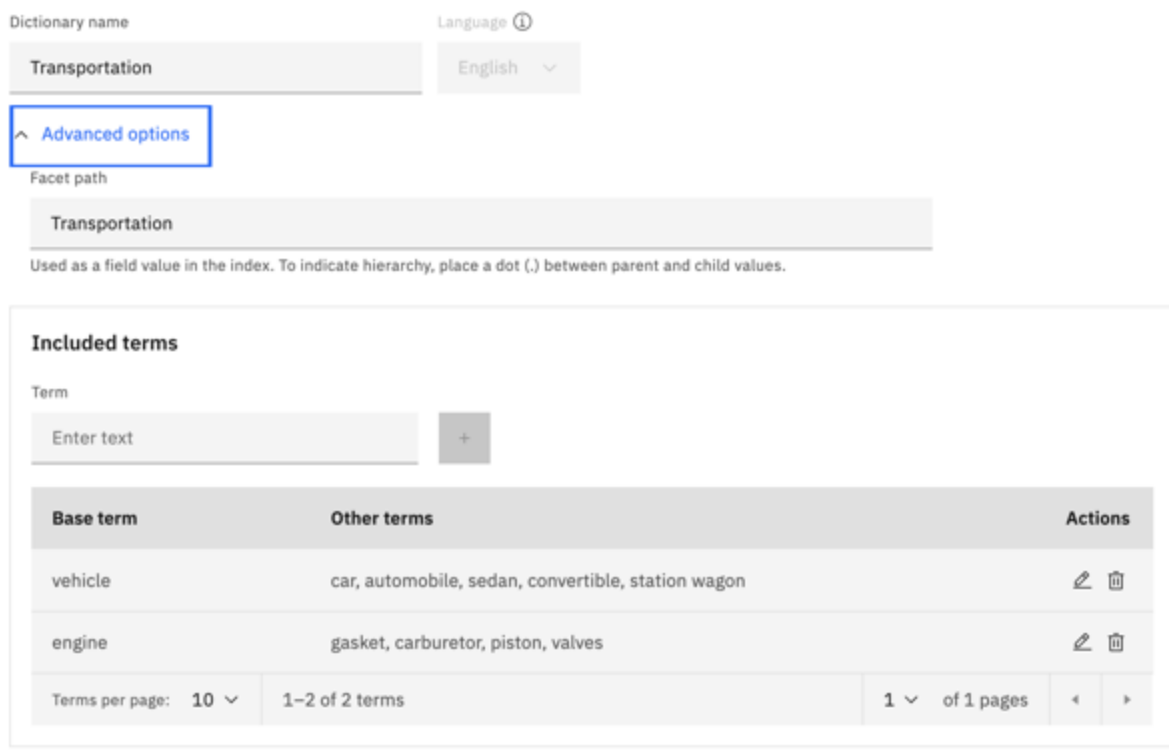

Figure 1. Transportation dictionary

The resulting facet that is created for the dictionary is displayed in the search page.

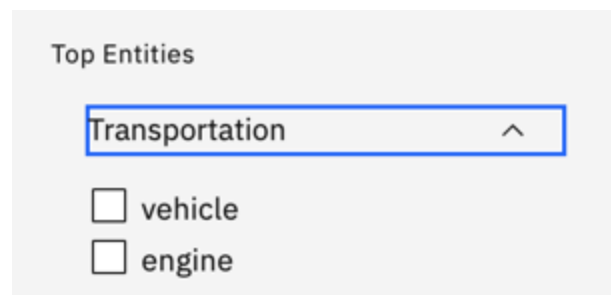

Figure 2. Transportation facet

The document where the enrichment is applied contains the following sentence:

```
Some car fluids can be acidic, such as battery fluid.
```

The following JSON snippet illustrates how a Transportation dictionary enrichment mention is stored when the term `car`, which is a synonym for the `vehicle` dictionary entry, is found in the document. In this collection, the dictionary enrichment is applied to the `text` field, so the mention is listed in the `entities` array that is in the `enriched_text` array.


```
{
   "enriched_text":  [
      {
         "entities":  [
            {
               "model_name":  "Dictionary:.Transportation",
               "mentions":  [
                  {
                     "confidence":  1,
                     "location":  {
                        "end":  91122,
                        "begin":  91119
                     },
                     "text":  "car"
                  }
               ],
               "text":  "vehicle",
               "type":  "Transportation"
            }
         ]
      }
   ]
}
```


Uploading dictionary terms

To add dictionary from a CSV file, complete the following steps:

1. Create a CSV file that contains the dictionary terms that you want to add.

   Use UTF-8 encoding. Specify one entry per line.

   - To define a set of synonymous terms, use the following syntax:

     ```
     <term>,<synonym>,<synonym>,<synonym>,...
     ```

     For example:

     ```
     vehicle,car,automobile,sedan,convertible,station wagon
     ```

     The entry in this example creates a `vehicle` dictionary entry. When the dictionary enrichment is applied to a document, any mentions of `vehicle`, `car`, `automobile`, `sedan`, `convertible`, or `station wagon` are tagged as instances of the `vehicle` dictionary entry.

   - To define a set of terms in the same category, use the following syntax:

     ```
     <category>,<related-term>,<related-term>,...
     ```

     For example:

     ```
     engine,gasket,carburetor,piston,valves
     ```

     The entry in this example creates an `engine` dictionary entry. When the dictionary enrichment is applied to a document, any mentions of `engine`, `gasket`, `carburetor`, `piston`, or `valves` are tagged as instances of the `engine` dictionary entry.

2. From the *Teach domain concepts* section of the *Improvement tools* panel, choose **Dictionaries**.
3. Click **Upload**.
4. Name your dictionary and choose the language that was used in the CSV file.
5. **Optional**: Expand *Advanced options*, and specify edit the facet name for the dictionary. Facets are used to categorize documents. A user can choose a facet type to narrow their search. The dictionary name in lowercase is used as the facet name by default. You might want to change the facet to be uppercase.
6. Click **Upload** to browse for the CSV file that you created earlier.
7. Click **Create**.

8. Choose the collections and fields where you want to apply the dictionary, and then click **Apply**.

> **Note:** If you add a dictionary by using the Enrichment API, after you apply the API-generated dictionary enrichment to a field, the dictionary is displayed in the Dictionaries page. However, you cannot edit the API-generated dictionary from the dictionary tool in the product user interface.

To delete a dictionary, you must use the Delete an enrichment method of the Discovery v2 API.

> **Note:** There is a limitation in how words with Hankaku (half-width) characters in Japanese are handled by the dictionary enrichment. When you create a dictionary enrichment in the Japanese language, you can use the Katakana or alphanumeric characters in the dictionary entry. However, when a Katakana word is used in the dictionary entry, the synonyms are handled with Zenkaku characters, except for the same Katakana word, which is represented by Hankaku characters. The Hankaku word is treated as a separate term from the Zentaku words. It is displayed as a separate facet, for example. Similarly, when an alphanumeric word is used in the dictionary entry, the synonyms are handled with Hankaku characters, except for the same alphanumeric word, which is represented by Zenkaku characters. The Zenkaku word is treated as a separate term from the Hankaku words.

Dictionary enrichments that you add to one project can be applied to collections in other projects in the same service instance. In fact, you can apply them to collections in a Content Mining project from the deployed Content Mining application.

Dictionary limits

The number of dictionaries and term entries you can create per service instance depends on your Discovery plan type.

Plan	Number of dictionaries per service instance	Number of term entries per dictionary	Number of terms for which suggestions can be generated
Cloud Pak for Data	Unlimited	Unlimited	1,000
Premium	100	10,000	1,000
Enterprise	100	10,000	1,000
Plus (includes Trial)	20	1,000	50

Dictionary plan limits

Define custom entities

Teach Discovery about terms that are significant to your business by creating an entity extractor.

An *entity extractor* is a machine learning model that recognizes and tags terms that you indicate are significant to your business need or use case. When you create an entity extractor, you get to decide the content and scope of information to find and extract. Your extractor can extract any of the following things:

- Terms that represent objects, such as vegetable names from cooking recipes or the make and model of cars from accident reports
- Attributes of objects, such as color and quantity
- Short phrases, such as `107 deaths in France`, `revenue of $343M`

An *entity type* is a type of thing. To create an entity extractor, you define a set of *entity types* that you care about. You then annotate a collection of your own documents by finding terms or phrases that represent the type of information you want to extract and labeling them as entity examples.

After you define entity types and label entity examples, you can generate a machine learning model. The model learns about the information that you care about based on how the terms or phrases that you label as examples are referenced in sentences. The model learns from the context and language with which the entity examples are referenced in the training data.

After the machine learning model is trained well enough to recognize your entity types, you can publish the model as an enrichment and apply the enrichment to new documents. The custom entity extractor enrichment recognizes and tags new mentions of the same and similar terms as occurrences of the entity types that you care about.

For more information about how to use the entity extractor to add domain customization to your AI applications, see the Entity Extractor Feature in Watson Discovery v2 blog post.

Discovery also has a built-in *Entities* enrichment that can be applied directly to your collection. It doesn't require any training to recognize commonly-known proper nouns. For more information about the Watson NLP Entities enrichment, see Entities.

You already built an entity type system in Knowledge Studio? You can use the corpus that is associated with your machine learning model as a starting

point for your entity extractor training data. For more information, see Importing a corpus.

For information about the languages with which the entity extractor can be used, see Language support.

Entity extractor overview video

This video provides an overview of how to define custom entity types and then use them to extract terms of interest from your data.

View video: Define custom entity types with Watson Discovery

To read a transcript of the video, open the video on YouTube.com, click the *More actions* icon, and then choose *Open transcript*.

Example

If you are familiar with the built-in Entities enrichment, you know that the enrichment can recognize terms that match generalized categories, such as `Person` and `Location`. With the entity extractor, you control what constitutes terms or phrases that are meaningful.

The following image shows the terms that an enrichment that recognizes `family members` entity type mentions might extract from text. The example illustrates how family member mentions and other entity mentions (that are recognized by the built-in Entities enrichment) both might be predicted.

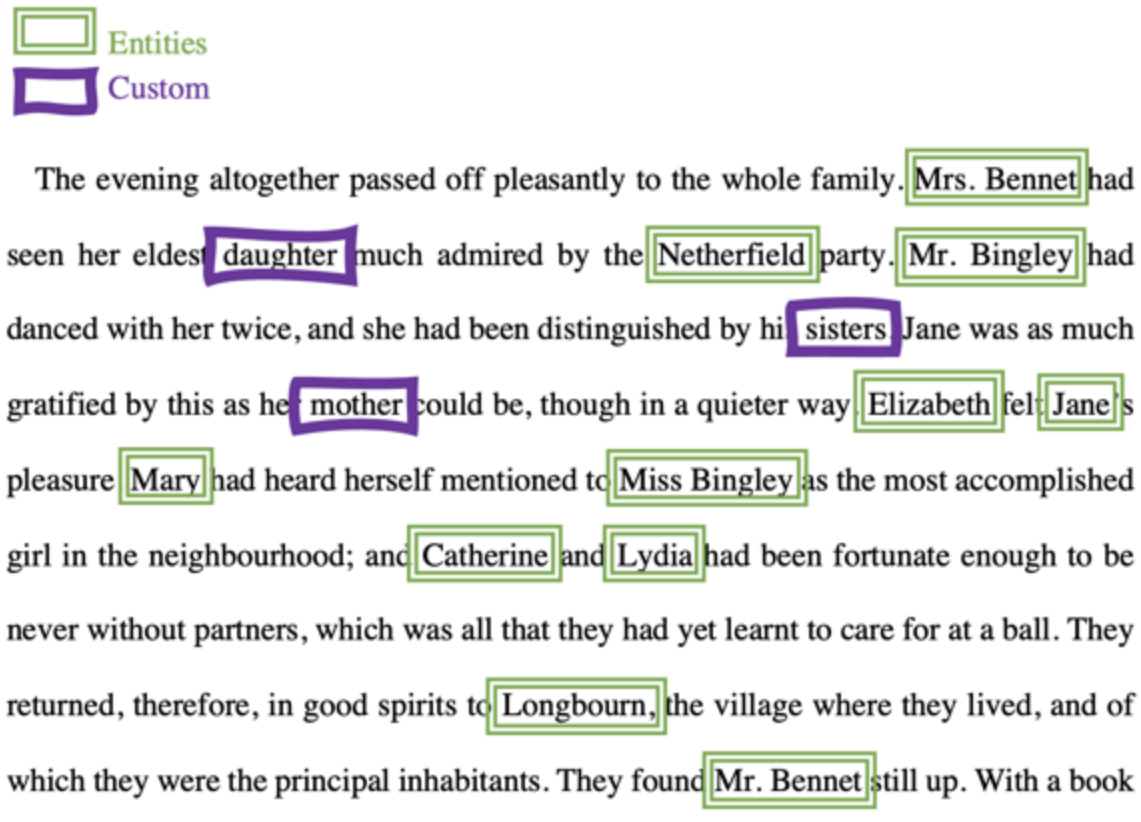

Figure 1. Labeled entity examples

This excerpt comes from Chapter 3 of *Pride and Prejudice* by Jane Austen.

Before you begin

Find or create a collection with documents that have various examples of the entity types that you want Discovery to learn about. To teach the extractor, you must label examples of entity types. You can only label examples if your collection contains valid examples. Try to find documents that have many and varying terms that function as examples of every entity type that you want to define.

Adding an entity extractor

To add an entity extractor, complete the following steps:

1. Open the project where you want to create the entity extractor.

   The project must have at least one collection with documents that are representative of your domain data.

2. From the *Improvement tools* panel of the *Improve and customize* page, expand **Teach domain concepts**, and then click **Extract entities**.

3. Click **New**.

   If you want to create an entity extractor that is based on the entity type system from a IBM Watson® Knowledge Studio corpus, click the arrow, and then choose **Import a Knowledge Studio corpus**. For next steps, see Importing a Knowledge Studio corpus.