



KeXtract

DOCUMENTAZIONE API

Versione	1.1
Stato:	RELEASED
Data:	20/01/2026

Kedos^{Srl} - Information Technology Solutions

DIVISIONI OPERATIVE

Viale Monza, 1 - 20125 Milano (MI)
C.so Inghilterra, 49 - 10138 Torino (TO)
Strada G. Inzani, 5 - 43125 Parma (PR)

SEDE LEGALE: via Chiavari, 5/E - 43125 Parma (PR),

P.IVA 02629980349, - R.I. PR-254346 Capitale sociale: 220.000,00€ i.v. - **Tel. 0236741057 - Fax 0521.995819**



Documentazione API KeXtract

Informazioni Generali

Base URL: `https://api.kextract.it/prod`

Autenticazione: Tutte le richieste richiedono l'header `x-api-key` con una chiave API valida.

1. Inizializzazione Documento

Endpoint

POST `/kextract/document`

Descrizione

Questo endpoint inizializza il processo di caricamento asincrono di un documento. Restituisce un URL pre-firmato per il caricamento del file sullo storage S3 di AWS (di seguito abbreviato S3) e un `document_id` univoco che identifica il documento.

Headers

Header	Tipo	Obbligatorio	Descrizione
<code>x-api-key</code>	string	Sì	Chiave API per l'autenticazione

Body Request

```
{
  "filename": "string"
}
```

Campo	Tipo	Obbligatorio	Descrizione
<code>filename</code>	string	Sì	Nome del file da caricare (inclusa l'estensione)

Esempio Request

```
{
  "filename": "Elenco Fornitori_Aggiornato 10ago2022.pdf"
}
```

Response (200 OK)

```
{
  "statusCode": 200,
  "body": {
    "presigned_url": "string",
    "document_id": "string"
  }
}
```

Campo	Tipo	Descrizione
<code>statusCode</code>	number	Codice di stato HTTP
<code>body.presigned_url</code>	string	URL pre-firmato S3 per il caricamento del file (valido per 1 ora)

body.document_id	string	ID univoco del documento (UUID)
------------------	--------	---------------------------------

Esempio Response

```
{
  "statusCode": 200,
  "body": {
    "presigned_url": "https://s3.eu-south-1.amazonaws.com/kextract-test-async/documents/78cf6536-9d55-45d7-bfdf-e26aefae9442/Elenco%20Fornitori_Aggiornato%2010ago2022.pdf?X-Amz-Algorithm=AWS4-HMAC-SHA256&...",
    "document_id": "78cf6536-9d55-45d7-bfdf-e26aefae9442"
  }
}
```

Note

- Il `presigned_url` ha una validità di 3600 secondi (1 ora)
- Il `document_id` deve essere conservato per le successive chiamate API
- Dopo aver ricevuto la response, è necessario caricare il file su S3 utilizzando l'URL pre-firmato

Prossimo Step

Utilizzare il `presigned_url` per caricare il file tramite una richiesta PUT. Vedere la sezione [Caricamento File su S3](#).

2. Caricamento File su S3

Endpoint

```
PUT {presigned_url}
```

Descrizione

Dopo aver ottenuto l'URL pre-firmato dall'endpoint `/kextract/document`, utilizzare questo URL per caricare il file direttamente su Amazon S3. Questa è una chiamata diretta a S3, non all'API KeXtract.

Headers

Header	Tipo	Obbligatorio	Descrizione
Content-Type	string	Sì	Tipo MIME del file (es: application/pdf, image/png, ecc.)

Body Request

Il body della richiesta deve contenere il file binario in formato `multipart/form-data`.

Esempio Request

```
PUT https://s3.eu-south-1.amazonaws.com/kextract-test-async/documents/78cf6536-9d55-45d7-bfdf-e26aefae9442/Elenco%20Fornitori_Aggiornato%2010ago2022.pdf?X-Amz-Algorithm=...
Content-Type: application/pdf

[File binario]
```

Response (200 OK)

In caso di successo, S3 restituisce una risposta con status code 200.

Note

- Utilizzare il `presigned_url` ottenuto dalla chiamata precedente

- L'URL è valido solo per 1 ora (3600 secondi)
- Il documento può avere una dimensione massima di 50MB
- Il documento PDF può avere al massimo 100 pagine
- Il Content-Type deve corrispondere al tipo di file caricato (pdf o immagine in formato PNG o JPG)
- Questa è una richiesta diretta a S3, non richiede l'header `x-api-key`

Prossimo Step

Dopo aver caricato il file su S3, procedere con l'avvio dell'elaborazione. Vedere la sezione [Avvio Elaborazione](#).

3. Avvio Elaborazione

Endpoint

POST /kextract/run

Descrizione

Avvia l'elaborazione asincrona del documento caricato su S3. Questo endpoint permette di specificare il tipo di estrazione, il formato di output e lo JSON Schema per l'estrazione strutturata dei dati.

Headers

Header	Tipo	Obbligatorio	Descrizione
x-api-key	string	Sì	Chiave API per l'autenticazione

Body Request

```
{
  "id_doc": "string",
  "media_type": "string",
  "extraction_type": "string",
  "output_type": "string",
  "zoom_quality": number,
  "schema": {}
}
```

Campo	Tipo	Obbligatorio	Descrizione
id_doc	string	Sì	ID del documento ottenuto dalla chiamata /kextract/document
media_type	string	Sì	Tipo MIME del documento (es: application/pdf)
extraction_type	string	Sì	Tipo di estrazione: query per estrazione strutturata oppure all per estrazione testuale
output_type	string	Sì	Formato di output: json per estrazione strutturata oppure Markdown o Plain Text per estrazione testuale
zoom_quality	number	No	Fattore di zoom per la qualità dell'elaborazione (possibili valori 1 oppure 2, opzione di default: 1). Parametro considerato solo per i documenti in formato PDF
schema	object	No	JSON Schema che definisce la struttura dei dati da estrarre. Parametro considerato solo se extraction_type è valorizzato a 'query'

JSON Schema

Lo schema segue la specifica JSON Schema Draft 2020-12. Definisce la struttura dei dati che si desidera estrarre dal documento.

Esempio Request

```
{
  "id_doc": "78cf6536-9d55-45d7-bfdf-e26aefae9442",
  "media_type": "application/pdf",
  "extraction_type": "query",
  "output_type": "json",
  "zoom_quality": 1,
  "schema": {
    "$schema": "https://json-schema.org/draft/2020-12/schema",
    "title": "Response",
    "type": "object",
    "properties": {
      "elenco_fornitori": {
        "type": "array",
        "description": "elenco_fornitori",
        "items": {
          "type": "object"
          "properties": {
            "fornitore": {
              "type": "string",
              "description": "fornitore"
            },
            "citta": {
              "type": "string",
              "description": "citta"
            },
            "nazione": {
              "type": "string",
              "description": "nazione"
            }
          }
        }
      },
      "required": []
    }
  },
  "required": []
}
```

Response (200 OK)

```
{
  "statusCode": 200,
  "body": {
    "elaboration_id": "string"
  }
}
```

Campo	Tipo	Descrizione
statusCode	number	Codice di stato HTTP
body.elaboration_id	string	ID univoco dell'elaborazione (UUID)

Esempio Response

```
{
  "statusCode": 200,
  "body": {
    "elaboration_id": "3d2ee6cb-55b7-4ae0-8a29-7e7c06ba3ab4"
  }
}
```

Note

- L' `elaboration_id` deve essere conservato per verificare lo stato dell'elaborazione
- L'elaborazione è asincrona e richiede tempo per essere completata.
- Lo JSON Schema definisce la struttura esatta dei dati che si vuole estrarre
- Il campo `zoom_quality` influenza la qualità dell'elaborazione dei documenti PDF derivanti da scansioni o comunque contenenti immagini di documenti (impostando il valore `2` viene applicato un upscaling che può consentire una maggiore precisione nella lettura dei testi ma può comportare un tempo di elaborazione più lungo e un maggiore consumo di token)

Prossimo Step

Utilizzare l' `elaboration_id` per verificare lo stato dell'elaborazione. Vedere la sezione [Verifica Stato Elaborazione](#).

4. Verifica Stato Elaborazione

Endpoint

```
GET /kextract/result/{elaboration_id}
```

Descrizione

Verifica lo stato dell'elaborazione asincrona. L'elaborazione richiede tempo e questa chiamata deve essere effettuata tramite polling fino al completamento.

Headers

Header	Tipo	Obbligatorio	Descrizione
x-api-key	string	Sì	Chiave API per l'autenticazione

URL Parameters

Parametro	Tipo	Obbligatorio	Descrizione
elaboration_id	string	Sì	ID dell'elaborazione ottenuto dalla chiamata <code>/kextract/run</code>

Esempio Request

```
GET /kextract/result/3d2ee6cb-55b7-4ae0-8a29-7e7c06ba3ab4
x-api-key: DS4*****apw
```

Response - Status IN_PROGRESS (200 OK)

Quando l'elaborazione è ancora in corso:

```
{
  "statusCode": 200,
  "body": {
    "startTime": "2025-12-10T15:19:43+00:00",
    "elabId": "3d2ee6cb-55b7-4ae0-8a29-7e7c06ba3ab4",
    "documentId": "78cf6536-9d55-45d7-bfdf-e26aefae9442",
    "elaborationType": "query",
  }
}
```

```

    "status": "IN_PROGRESS"
  }
}

```

Campo	Tipo	Descrizione
statusCode	number	Codice di stato HTTP
body.startTime	string	Data e ora di inizio elaborazione (formato ISO 8601)
body.elabId	string	ID dell'elaborazione
body.documentId	string	ID del documento
body.elaborationType	string	Tipo di elaborazione
body.status	string	Stato corrente: IN_PROGRESS

Response - Status COMPLETED (200 OK)

Quando l'elaborazione è completata:

```

{
  "statusCode": 200,
  "body": {
    "error_message": "",
    "key_s3": "documents/3d2ee6cb-55b7-4ae0-8a29-7e7c06ba3ab4.json",
    "documentId": "78cf6536-9d55-45d7-bfdf-e26aefae9442",
    "compile_time": 58.32548322143555,
    "status": "COMPLETED",
    "zoom_factor": 1,
    "startTime": "2025-12-10T15:19:43+00:00",
    "zero_ret": false,
    "endTime": "2025-12-10T15:21:10+00:00",
    "token_in": 24520,
    "token_out": 4000,
    "elabId": "3d2ee6cb-55b7-4ae0-8a29-7e7c06ba3ab4",
    "elaborationType": "query",
    "presigned_url": "https://s3.eu-south-1.amazonaws.com/hsa-hzo/documents/3d2ee6cb-55b7-4ae0-8a29-7e7c06ba3ab4.json?X-Amz-Algorithm=..."
  }
}

```

Campo	Tipo	Descrizione
statusCode	number	Codice di stato HTTP
body.status	string	Stato finale: COMPLETED
body.elabId	string	ID dell'elaborazione
body.documentId	string	ID del documento
body.elaborationType	string	Tipo di elaborazione
body.startTime	string	Data e ora di inizio elaborazione (formato ISO 8601)
body.endTime	string	Data e ora di fine elaborazione (formato ISO 8601)
body.compile_time	number	Tempo di elaborazione interno in secondi
body.key_s3	string	Chiave S3 del file risultato

body.presigned_url	string	URL pre-firmato per scaricare il risultato (valido 30 minuti)
body.token_in	number	Numero di token in input utilizzati
body.token_out	number	Numero di token in output generati
body.zoom_factor	number	Fattore di zoom utilizzato
body.zero_ret	boolean	Indica la modalità Zero Retention
body.error_message	string	Messaggio di errore (vuoto se non ci sono errori)

Note

- **Strategia di Polling:** È consigliabile effettuare la prima chiamata dopo **30 secondi** dall'avvio dell'elaborazione
- **Intervallo minimo:** Non effettuare chiamate con intervallo inferiore a **10 secondi**
- **Timeout:** Se non si riceve risposta entro **15 minuti**, la chiamata può considerarsi persa
- Il `presigned_url` nella response COMPLETED è valido per **1800 secondi (30 minuti)**
- Utilizzare il `presigned_url` per scaricare i risultati dell'estrazione

Stati Possibili

- `IN_PROGRESS` : Elaborazione in corso
- `COMPLETED` : Elaborazione completata con successo
- `ERROR` : Elaborazione completata con errore nb: in caso di errore, il campo `error_message` contiene informazioni dettagliate sull'errore.

Prossimo Step

Quando lo stato è `COMPLETED`, utilizzare il `presigned_url` per scaricare il file JSON con i risultati dell'estrazione. Vedere la sezione [Download Risultati](#).

5. Download Risultati

Endpoint

```
GET {presigned_url}
```

Descrizione

Scarica il risultato dell'elaborazione, utilizzando l'URL pre-firmato ottenuto dalla risposta COMPLETED della verifica dello stato. Questa è una chiamata diretta a S3 in sola lettura.

Headers

Nessun header particolare richiesto. L'autenticazione è gestita tramite i parametri dell'URL pre-firmato.

Esempio Request

```
GET https://s3.eu-south-1.amazonaws.com/kextract-test-async/documents/3d2ee6cb-55b7-4ae0-8a29-7e7c06ba3ab4.json?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=...&X-Amz-Signature=...
```

Response (200 OK)

Il contenuto della response va interpretato in accordo con quanto specificato nel parametro `output_type`

Esempi Response

In caso di `output_type` valorizzato `JSON`

```
{"elenco_fornitori": [{"fornitore": "Nome Fornitore", "citta": "Milano", "nazione": "Italia"}, {"fornitore":
```

```
"Altro Fornitore", "citta": "Roma", "nazione": "Italia"]}]}
```

In caso di `output_type` valorizzato `Markdown`

```
# calendario trattamenti BIO

## POMACEE
**dosaggi x 10 litri di acqua**

| | | | | dosaggio | carotene |
|-----|-----|-----|-----|-----|-----|
| ottobre | ½ caduta foglie | nectria | ossicloruro rame | 60 gr | 20 gg |
| novembre | al bruno | nectria | solfato rame | 80 gr | 20 gg |
| dicembre | al bruno | nectria | poltiglia bordolese | 100 gr | 20 gg |
...
| luglio | 3 volte | carpocapsa | samosad+bacillus thuring | vedi etichette | |
```

In caso di `output_type` valorizzato `Plain Text`

```
calendario trattamenti BIO

POMACEE
dosaggi x 10 litri di acqua

dosaggio carotene
ottobre ½ caduta foglie nectria ossicloruro rame 60 gr 20 gg
novembre al bruno nectria solfato rame 80 gr 20 gg
dicembre al bruno nectria poltiglia bordolese 100 gr 20 gg
...
luglio 3 volte carpocapsa samosad+bacillus thuring vedi etichette
```

Note

- Questa è una chiamata diretta a S3, non all'API KeXtract
- Non richiede l'header `x-api-key`
- L'URL pre-firmato è valido per **30 minuti (1800 secondi)**
- Il file può essere scaricato più volte finché l'URL pre-firmato è valido

Flusso Operativo Completo