European Court of Auditors Corpus

M.T. Carrasco Benitez, 13 April 2026

Abstract

The *European Court of Auditors Corpus* (ECA Corpus) is a system for public documents that the ECA may designate as part of its corpus, if and when such a corpus is defined or revised. The corpus is intended to represent the ECA acquis, publications such as special reports and formal documents such as speeches by ECA presidents. It excludes internal documents and documents about the ECA produced by other entities, such as peer reviews. The IT side of the project has been completed; the **next step** must be to create a complete *Metadata Table* for the ECA Corpus.

Overview

The ECA Corpus is a **registry** in which each record (called *dossier*) contains both metadata and full-text. The ECA Corpus is provided as a *Portable Web Package*, a simple directory structure that can be opened with web browsers without installation; it contains only data and does not include programs.

It is generated from the Metadata Table and is available online at *ecacorpus.eu*, or can be downloaded for local use. To explore it locally, simply unzip the package and open the *index.html* file in the root directory. The online and downloaded versions are identical, as long as they share the same timestamp.

http://ecacorpus.eu
https://ecacorpus.eu/misc/ecacorpus.zip
https://dragoman.org/xdossier/pwp.html

Next step

To support data completion and cleaning, all metadata is currently consolidated in a single Metadata Table, like a spreadsheet; to begin with, full-text is not taken into consideration. Details in:
https://ecacorpus.eu/metadata/index.html

The first step should be to export the metadata directly from the ECA systems, as this is the most effective and streamlined approach. If this is not possible, to edit the current Metadata Table. The primary fields to address initially are: *number*, *date*, *type*, *title*, and *ruri*. Full description in:
https://ecacorpus.eu/metadata/field.html

The *number* is currently implicit, determined by its position in the Metadata Table rather than being stated explicitly; once the table is finalised, the number will be fixed and recorded explicitly. Storing the data in a machine-processable format offers a significant advantage, as demonstrated by the generation of *ecacorpus*.

The metadata could be expanded, starting from the filters available in the ECA search interface, such as *Topic* and *Country covered*. Further metadata could also be extracted directly from the publications themselves, including details about the audit team, such a responsible Members and other contributors. In addition, new fields could be introduced to enhance interconnectivity, such as links to related dossiers or associated reports.

Metadata

Collecting the metadata was difficult because the ECA does not maintain a complete register of its publications. There are no machine-readable sources, only PDFs and a website without a public

API; access via Microsoft SharePoint is limited and inconsistent, making it hard to build a reliable Metadata Table. Instead, ECA recommended using its website search engine and five PDF lists: Annual Reports, Specific Annual Reports, Special Reports, Opinions, and Reviews:
https://ecacorpus.eu/misc/note.html

The current version of the Metadata Table has been compiled primarily by scraping the ECA search engine, reconciling the results with the five PDF lists, and cross-checking against the Publications Office and internal sources; in addition, certain data on the ECA website cannot be cleanly extracted from its HTML structure. Because this process is not fully reliable, the Metadata Table likely contains errors, and some publications may be missing; moreover, verification is not possible in the absence of a comprehensive register.

Despite these limitations, the exercise is worthwhile: the resulting Metadata Table is likely the most complete and structured dataset of ECA publications currently available, although further cleaning and refinement are still needed. The approach is highly empirical and reflects the practical realities of how institutions operate.

Cleaning
"§" refer to the dossier number (first column) in:
https://ecacorpus.eu/report/full.html

* Tasks:
- Add missing data since April 2025 § 2658
- Delete dossiers that should not be part of the corpus
- Add missing dossiers that should be included in the corpus
- Add English titles with links § 11
- Add title links § 4

* Link rules:
- Title links should point to the English ECA landing page § 2658
- Official Journal (OJ) links should point to EUR-Lex, which provides multilingual versions and formats
- If only the OJ version is available on the ECA website, link instead to the English EUR-Lex PDF, which is usually of better quality than the ECA version § 10

* Date
Which date should be indicated? Example with Annual report of 1977 § 10:
- ECA website: 01-01-1978
- ECA: 30 November 1978
- OJ: 30 December 1978

Dossier types
Each dossier has one *type*, labelled by one mnemonic word: *budget*, *specific*, *info*, *opinion*, *paper*, *work*, *activity*, *speech*, *preview*, *journal*, *decision*, and *misc*. Full description in:
https://ecacorpus.eu/misc/type.html

The Annual Report files for some years exceed 1,000 which correspond to the types: *budget*, *specific*, and *info*.

Repository
The *repository folder* contains folders for each dossier, named by their dossier number. It is currently a nearly empty structure prepared to store full-text documents in multiple languages and formats, starting with English:
https://ecacorpus.eu/repository

Each dossier folder is also designed as a PWP and will eventually include its own metadata. In the future, the Metadata Table will be generated from the metadata stored within these folders, but for now, work is carried out directly in the central Metadata Table for practical reasons.

Similar to the handling of metadata, the most effective and streamlined approach is to export content directly from the ECA systems, which are believed to be based on Microsoft SharePoint. This export should be delivered in a standard, structured format that is independent of Microsoft-specific tools. It should include both metadata and the full-text of all documents, with PDFs converted into plain UTF-8 text to ensure ease of processing and long-term accessibility.

When a document contains multiple sections, only the relevant parts should be included. For example, in OJ C139/1979 (§ 11), only the six opinions starting on page 15 that relate to the ECA should be placed in the corresponding dossier folder. All content should be converted to plain text to support further processing and ensure long-term preservation.

Dossier folder
Each dossier, including both metadata and full text, will eventually be stored in a folder named after its dossier number, structured as a PWP. Example and illustration:
https://ecacorpus.eu/repository/1/index.html
https://ecacorpus.eu/misc/dossier.html

General case
The project aims to show that a PWP can store a large collection of institutional documents, like the ECA corpus. This approach is simple, relying on web browsers for local or server access, yet robust and suitable for **long-term preservation**. The data can be accessed directly from the file system for traditional analysis or modern methods such as AI, and it can also be converted into a database system like SQLite.

The ECA Corpus is a specific example. The ECA was chosen because it is a European institution with a manageable volume of publications, making it feasible to build a complete corpus. Having all the data easily accessible supports more detailed analysis and research.

Inquires to ECA
Following email inquiries to the ECA, it was confirmed that:

- The ECA does not maintain a complete list of publications since 1977; only the website search engine and a few document lists are available.

- Most decisions are for internal use, so there is no publicly accessible list.

- The ECA publishes several annual reports some year, even though TEU Article 287 refers to *annual report* in the singular.

- Documents such as Audit in Brief, FAQs, and the glossary are not considered part of the official

annual reports.

Notices

- Disclaimer

Independent project not officially related to other parties such as the European Court of Auditors (ECA) or the Publications Office of the European Union (PO). For official sources check with the relevant parties:
https://eca.europa.eu
https://op.europa.eu

- Acknowledgment

Main data sources: ECA, PO.
The author thanks the ECA-INFO team that patiently answered the email inquiries.

- License

- Author

Manuel Tomas Carrasco Benitez
https://dragoman.org/carrasco
mtcarrasco@gmail.com

- Living document

The latest version of this document in:
https://ecacorpus.eu/misc/corpus.pdf

seco y sin flover

Ω