European Court of Auditors Corpus

M.T. Carrasco Benitez, 22 June 2024

Abstract

The *European Court of Auditors Corpus* (ECA Corpus) is a system for the public documents that could be considered the *ECA acquis*: publications such as *special reports*, or formal documents such as ECA presidents speeches. It excludes internal documents or documents about ECA produced by other entities such as peer reviews.

Package and visualisation

The *Corpus Dossier* is an interlinked web-based package adapted to web browsers and servers that contains known ECA publications since its inception in 1977, with internal links within the Dossier and to external resources such as ECA or Publications Office, particularly for the full data in several formats (PDF, HMTL, JPEG, etc) and languages (English, French, etc). The intention is to have a reliable source, facilitate the exploration, easy comparison for aspects such as publications per year of different types, or title naming patterns. A Dossier is about 10MB.

System

Each publication is a dataset containing metadata and data. The metadata contains elements such as title, date, type, links to data, etc. Though the data in several formats and languages is hosted externally, there can also be local copies in the Dossier. A Dossier with all the data should a few gigabytes.

The Dossier is created from the metadata and local copies, if available. The Dossier is also designed for long-term digital preservation as it does not require specific programs. Data formats should be viewable with web browsers without additional modules and they must have a good sporting chance of being processable in 100 years; hence, proprietary formats should be avoided and if in doubt, other versions in more reliable formats should be added.

Data

Creating a database with the metadata is challenging as ECA confirmed that it does not have a comprehensive register, the recommendation was to use the online search engine and links to facilitated publication lists, although there are discrepancies; for example, the list of annuals reports contains 46, the search engine returns 70 and the database in this project already has 820.

The online search engine does not provide an interface for downloading data. Hence, data comes mainly from (unreliably) scraping the search engine, comparing with the provided lists, cross-checking with the Publications Office and internal checks; so the database probably contains errors and publications might be missing as it is impossible to know with certainty the list of publications due to the lack of a register. Some data in HTML in the ECA site cannot be cleanly downloaded.

Even with all these shortcomings, this data governance exercise is worthwhile: the resulting database is arguably the most complete and structured dataset of the ECA publications, though further cleaning and refinements are needed. The whole approach is vey empirical and it shows the realities of a working institution.

General application

ECA was selected for this project because it is a European institution and the production of publications is of reasonable size; hence, constructing a full corpus is feasible in a project of this magnitude. Starting from the particular case of ECA, the theoretical outcomes are general and

applicable to similar organisations: it is a very humbling and sobering exercise to construct this type of database and downloading the data from a working organisation; one should expect similar situations with similar organisations.

Requests to ECA
Following emails requests to ECA, the following was established:

- ECA does not have a list of all publications since 1977. The only available sources are the website search engine and lists for some documents types.

- ECA does not have a publicly available list of decisions as the majority is only for internal use.

- ECA prepares several annual reports per year, though TEU article 287 mentions *annual report* in singular.

- Audit in brief, most frequently asked questions, and glossary are not part of annual reports proper.

To do
- Establish a list of all annual reports (ECA list: 46 | search engine English: 70 | Corpus: 820)
- Data cleaning
- Expand metadata: subject, countries, link publications, etc.
- URIs for data downloading
- Download data
- Data mining
- Comparison of multilingual parallel files

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.

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

- License

- Author
Manuel Tomas Carrasco Benitez
mtcarrasco@gmail.com

- Status
Ongoing project. Living document, latest version in `http://ecac.site/deco.pdf`