Documentation: Indexex

By Jeúsa Hamer

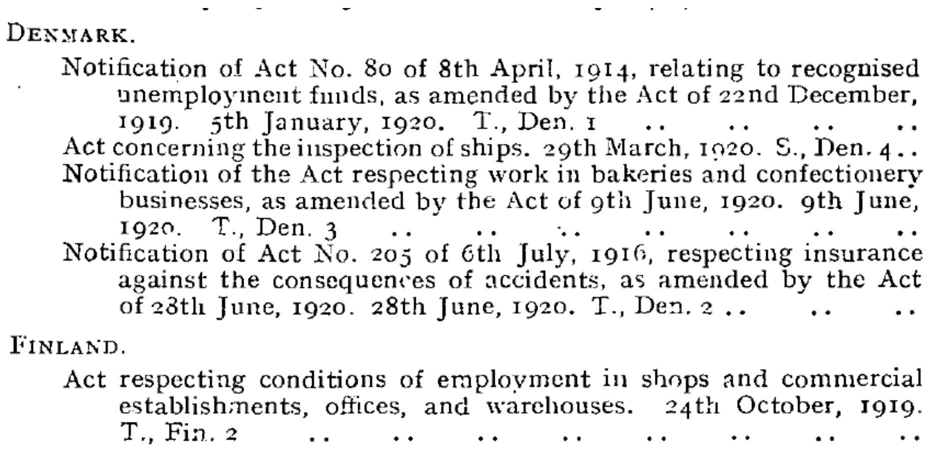
Introduction

Indexex is a command line application that extracts information from documents published by the International Labour Organisation containing an index of laws and orders passed by different countries. For every document, a table is generated, which is saved in a CSV-file. The extracted information entails short descriptions of the laws, the country that passed them and when they where passed.

Indexex can be used in two different modes: TESS and FITZ. FITZ means that the existing optical character recognition (OCR) of a PDF-file is used to read the text. TESS uses the Tesseract-OCR-engine to generate new OCR for the document.

The application should also be able to recognize and extract the index from a double paged document, but the extraction only works in mode TESS.

General functionality



DENMARK.

Notification of Act No. 80 of 8th April, 1914, relating to recognised unemployment funds, as amended by the Act of 22nd December, 1919. 5th January, 1920. T., Den. 1 .. ..

Act concerning the inspection of ships. 29th March, 1920. S., Den. 4 ..

Notification of the Act respecting work in bakeries and confectionery businesses, as amended by the Act of 9th June, 1920. 9th June, 1920. T., Den. 3 .. ..

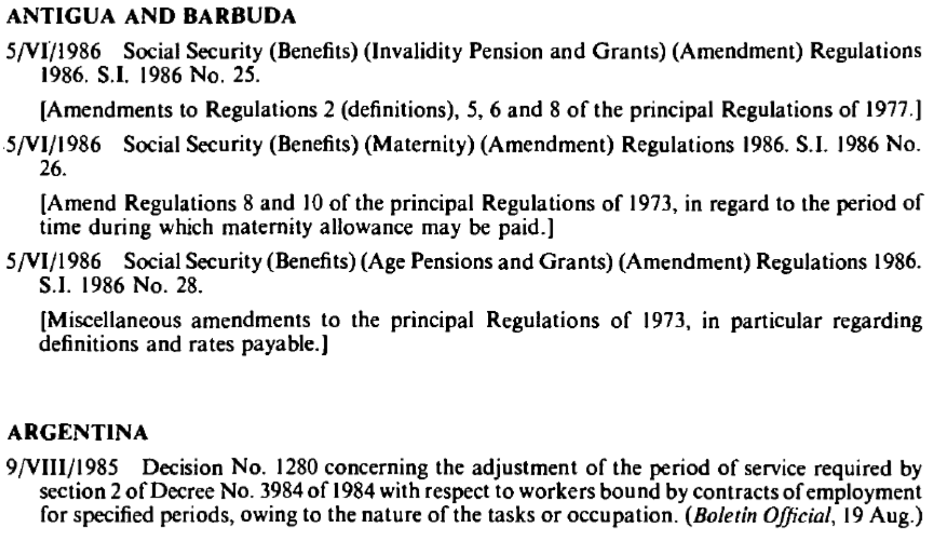
Notification of Act No. 205 of 6th July, 1916, respecting insurance against the consequences of accidents, as amended by the Act of 28th June, 1920. 28th June, 1920. T., Den. 2 .. .. ..

FINLAND.

Act respecting conditions of employment in shops and commercial establishments, offices, and warehouses. 24th October, 1919. T., Fin. 2 .. ..

Figure 1: Index from 1920

For the extraction of the countries and the laws from these documents, the indentations of the lines play a significant role. The laws are always sorted by country, meaning there is a line with the name of a country followed by the laws that were passed by this country. The country names, the first line of an entry and the following lines of an entry all have different indentation as you can see in the example in Figure 1. In this index there are three types of indentation. Therefore, this information can be used to extract the individual entries and the country names. If a line starts on the very left, it gets the label “country”. If a line is indented a bit, it gets the label “start”. The entries can then be extracted by taking a “start” line and its following lines until another line of type “start” or “country” is reached. The country that should be assigend to an entry can be found in the previous line labeled “country”.



ANTIGUA AND BARBUDA

5/VI/1986 Social Security (Benefits) (Invalidity Pension and Grants) (Amendment) Regulations 1986. S.I. 1986 No. 25.

[Amendments to Regulations 2 (definitions), 5, 6 and 8 of the principal Regulations of 1977.]

5/VI/1986 Social Security (Benefits) (Maternity) (Amendment) Regulations 1986. S.I. 1986 No. 26.

[Amend Regulations 8 and 10 of the principal Regulations of 1973, in regard to the period of time during which maternity allowance may be paid.]

5/VI/1986 Social Security (Benefits) (Age Pensions and Grants) (Amendment) Regulations 1986. S.I. 1986 No. 28.

[Miscellaneous amendments to the principal Regulations of 1973, in particular regarding definitions and rates payable.]

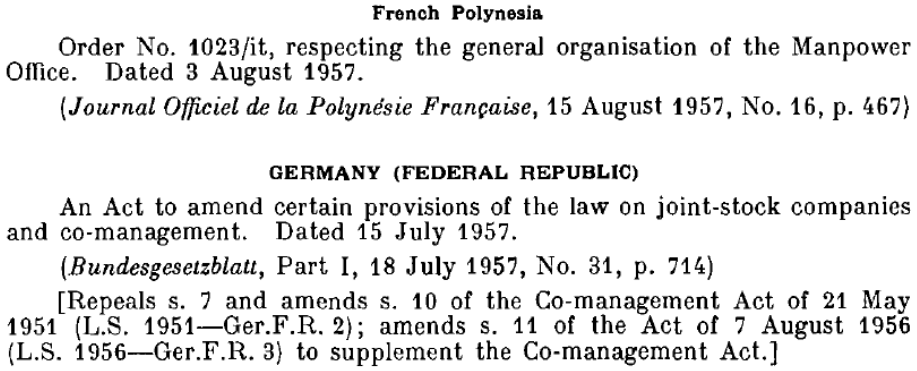
ARGENTINA

9/VIII/1985 Decision No. 1280 concerning the adjustment of the period of service required by section 2 of Decree No. 3984 of 1984 with respect to workers bound by contracts of employment for specified periods, owing to the nature of the tasks or occupation. (*Boletin Official*, 19 Aug.)

Figure 2: Index from 1986

Other documents, like the one in Figure 2, have another format with a different usage of indentation. Here, there are only two types of indentation and both the countries and the "start" lines start by the left text border. In these documents, where the lines end is also taken into account. Lines of type "start" usually end by the right text border, whereas "country" lines are much shorter. This information is used to distinguish the two types.

The assignment of these labels is done in the script `label.py`, especially in the method `assign_labels`. For this to work for these two different types of formats, in a previous part of the program it is determined if there are three or two identation types in the respective document.



French Polynesia

Order No. 1023/it, respecting the general organisation of the Manpower Office. Dated 3 August 1957.

(*Journal Officiel de la Polynésie Française*, 15 August 1957, No. 16, p. 467)

GERMANY (FEDERAL REPUBLIC)

An Act to amend certain provisions of the law on joint-stock companies and co-management. Dated 15 July 1957.

(*Bundesgesetzblatt*, Part I, 18 July 1957, No. 31, p. 714)

[Repeals s. 7 and amends s. 10 of the Co-management Act of 21 May 1951 (L.S. 1951—Ger.F.R. 2); amends s. 11 of the Act of 7 August 1956 (L.S. 1956—Ger.F.R. 3) to supplement the Co-management Act.]

Figure 3: Index from 1958

Figure 3 shows a third way of formatting, which is very different from the rest. Only few documents use this way of identation, which is why Indexex was not developed to regcognize this type automatically. Still, the entries can be extracted from these types of documents, where the countries are centered and the “start” lines are indented, if the user sets the corresponding arguments when using the application.

On top of the country, the program also tries to identify the “region” of an entry. Some documents specify not only the country that passed a law, but also in which part of a country it was passed. Lines that were previously labeled as “country” get the new label “region” if most of the letters in this line are lower case since country names are usually written in only upper case letters.

Grouping lines based on indentation

To assign labels to the lines, the lines first need to be assigned to groups based on where they start, which is defined by the x0-coordinate of a line, and where they end, defined by its x1-coordinate. For a document like the one in Figure 1, all “country” lines should be in one group and all “start” lines should be in another group based on their respective x0-coordinate. This grouping mechanism is an essential part of the program and is done in `group.py` in the method `group_rows`.

The algorithm iterates through the lines from the beginning of a page to the end and creates groups containing lines with similar x0-coordinates while doing that. It does this individually for each page of the document. If a line fits into an existing group, meaning there are lines with similar x0-coordinates in it, it is added to that group. If it does not, a new group is created and the line is added to the new group. To determine whether a line fits into a group, the mean value for the x0-coordinates of the last two lines that have been added to the group is compared to the x0-coordinate of the current line. If the difference is smaller or equal to a specified maximum distance value, the current line is added to that group. Since only the last two lines of a group are looked at, the algorithm works fairly well with documents that where scanned a bit skewedly, e.g. the x0-coordinates of the lines in a group slowly become larger from the start to the end of a page. Depending on how many groups are created for the pages, one can deduce if there are two or three identation types for this document.

Date Extraction

When the countries and the descriptions of the laws have been extracted, the date is then extracted from the descriptions. The documents have different formats for dates and they can be in the beginning or the end of an entry. Regular expressions are used to extract the dates and to identify the date format. This happens in the sript `date.py`. All dates are transformed to the same format after extraction, which is d.m.. This works when the months have english names, and it should also work for spanish and french. The year is extracted from the date of the entry if possible. Otherwise the year is taken from the name of the file.

Entries where no date can be extracted can usually be sorted out since often they are not actually a law or an order but for example a footnote that was extracted based on its identation. The application removes entries where no date could be extracted by default, but there is an argument that can be used to keep all entries that were found.

Problems

Since the x0-coordinates of the lines are so important for the application to work, it can be quite problematic if there are artifacts on a page before the line actually starts due to bad scanning quality. They can lead to wrong x0-coordinates for a line when the OCR identifies them as characters, which messes with the algorithm. The Tesseract-OCR-engine seems to be especially bad at identifying artifacts and they are often recognized as characters. Indexex tries to identify artifacts in the beginning of a line and deal with them, but this only works alright if there are only few artifacts on a page. Tesseract also does not handle pages well that were scanned too skewedly. It leads to very bad word recognition, and pages like that also might pose a problem for the algorithm of Indexex even though it was designed to deal with slightly skewed pages.

The indentation based grouping of the lines often does not work on the first page of a document where the actual index starts and the extraction should begin. On these pages, there are usually some paragraphs preceding the list of laws. Since a page like that does not follow the expected format of the document, the entries on that page often cannot be extracted. Title pages on the other hand are mostly ignored and do not pose a problem.

Another problem concerning the OCR is in what order the words are read. For the algorithm to work, the text needs to be read line by line. If the mode TESS is chosen, it does this automatically. OCR-engines that try to recognize the formatting of a text automatically sometimes read the text column-wise, e.g. one column with the dates and one with the law descriptions. If that is the case, Indexex will produce wrong results.

For documents where the quality of the scan is not great, upper and lower case letters are not always identified very well by the OCR-engine. Therefore, mistakes can happen regarding the distinction between “country” and “region”. Moreover, in some documents bilateral treaties between two countries are identified as a “region” because they are written in mostly lower case letters, and the corresponding “country” that is extracted is then usually “INTERNATIONAL”.

Bad OCR can also lead to problems concerning the date extraction. Even though the regular expressions are made to deal with common mistakes that can happen, e.g. a “2” is recognized as a “Z”, digits or months that were read incorrectly can still lead to wrong dates. It can also lead to the extraction not working at all when a date cannot be recognized in the text because of bad OCR. Since entries where no date could be extracted are removed by default, it may happen that valid entries are also removed.

The extracted country and region names are the ones that were found in the text and are not normalized in any way. Therefore, different ways of spelling for a country may exist in the extracted indexes. If the OCR did not recognize the text correctly, there also might be misspelled names. The same applies to the law descriptions. These also can contain wrongly spelled words since Indexex does not correct them.

Description of the output table:

“country”: name of the country that passed the law or order
e.g. “AUSTRALIA”

“region”: if given, name of specific region or colony
e.g. “Queensland”

“text”: the description of the law, “full_text” minus the “extracted_date” and the text is cleaned a bit
e.g. “An Act to provide for the licensing...”

“date”: extracted date in format d.m.
e.g. “22.4.”

“year”: extracted year, either taken from “extracted_date” or from the name of the file
e.g. “1983”

“page”: page number the entry was found on
e.g. “2”

“extracted_date”: date extracted from “full_text”,
e.g. “22/IV/1983”

“full_text”: full description of the law with date and without cleaning
e.g. “22/IV/1983 An Act to provide for the licensing...”