Bremen, February 2021

Gabriella Skitalinskaya
Nils Düpont

OCR Report

WeSIS — Technical papers
No. 7

Global Dynamics of Social Policy CRC 1342

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Gabriella Skitalinskaya, Nils Düpont
OCR Report
SFB 1342 Technical Paper Series, 7
Bremen: SFB 1342, 2021

in alphabetical order
Nils Düpont 0000-0002-4766-9540
Gabriella Skitalinskaya 0000-0001-5006-6196

University of Bremen

SFB 1342 Globale Entwicklungsdynamiken von Sozialpolitik /
CRC 1342 Global Dynamics of Social Policy

Postadresse / Postaddress:
Postfach 33 04 40, D - 28334 Bremen

Website:
https://www.socialpolicydynamics.de

[DOI https://doi.org/10.26092/elib/1517]
[ISSN 2700-0389]

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG)
Projektnummer 374666841 – SFB 1342

Gabriella Skitalinska
Nils Düpont

OCR Report

SFB 1342
No. 7

OCR Report

Gabriella Skitalinskaya, Nils Düpont

Index

Overview and Recommendations

Based on sample files of the first co-creation workshop on "Optical Character Recognition (OCR)" on October 30th, 2018, we compare the results of different tools that are available for OCR as of April 2019. Our comparison of optical character recognition software includes:

- OCR engines that do the actual character identification
- Layout analysis software that divides scanned documents into zones suitable for OCR
- Graphical interfaces to one or more OCR engines

Depending on the capability of each solution, the tools had to perform three tasks:

- Extract plain text
- Recognize the text style and its structure (hOCR)
- Extract tables incl. the proper layout and corresponding cell data

In general, OCR remains a challenge in computer science and requires a huge amount of training algorithms to achieve sufficient accuracy. While some packages claim to achieve between 97% and 99% accuracy, note that these rates are based on character errors, not word errors.

The tools that were tested are Google Docs OCR, Tesseract, ABBYY FineReader, PDF OCR X, and Adobe Acrobat. They differ widely ranging from command-line tools to ones with advanced graphical user interfaces; some include the option to preprocess images, others not; likewise, some are free and open-source tools, some are commercial software. Depending on the task at hand, different tools achieve satisfiable results: for plain text extraction Adobe Acrobat performed quite well. While it is easy to apply, the advantage is that Acrobat is already installed on many CRC member's working machines. Regarding hOCR, ABBYY FineReader outperforms other tools. The same is true for extracting tables. Being more than 20 years on the market, ABBYY is the most sophisticated and advanced OCR tool offering the most flexibility. Yet, it is also the most expensive one.

Tesseract as a free and open-source tool may perform sufficiently well if the projects are able to control the workflow right from the beginning. If projects plan on digitizing sources and even can control the scanning thereby ensuring a high quality of the source material, Tesseract may be a choice as well.

As a general rule, the better the quality of the source material the better the OCR results. Projects are therefore well advised to acquire the best source material as possible and to clearly define the objective and desired output of their OCR task. Depending on the goal, the amount of material, the standardization and repeatability of the task, projects are also well advised to see if OCR is worth the effort at all – or if the same goal could be achieved by human codings as well. As OCR requires a huge amount of resources for training the algorithms to achieve accuracy and quality, A01 cannot contribute to the development of new OCR engines or algorithms *per se* but can help in finding a proper solution for a project's tasks and assist in setting up a workflow.

1. Introduction

On October 30, 2018, the first co-creation session on "Optical Character Recognition" (OCR) took place. During the workshop three points became clear:

1. Vanilla text OCR is not considered a major problem. Projects already have achieved satisfying results for plain text extraction with the tools they have available (i.e. Adobe Acrobat).
2. Extracting figures from PDFs is of interest.
3. Extracting tables is the main issue with which the projects are struggling.

OCR in general is an emerging field in computer science and numerous tools are already available. Most of them are commercial tools especially geared towards OCR (e.g. ABBY FineReader), others are "in-built" functions of broader applications (e.g. Adobe Acrobat). Recently, there has been increased research and interest in open-source solutions (e.g. tesseract). All tools differ widely regarding their main objective, their flexibility and their user-friendliness ranging from simple but powerful command-line tools to advanced graphical user-interfaces (GUI).

As OCR requires a large amount of training of the algorithm(s) with many images of each character, simple text recognition has matured to a high degree of accuracy. The difficulty, however, still lies in identifying the structure of a text (font size, captions, two-column layouts etc.) or recognizing tables. Furthermore, the success of an OCR heavily depends on the quality of the source material. For this reason, some tools apply preprocessing techniques to increase picture quality, some can be used to build a workflow around them. Yet, they cannot add what is already lost when having distorted images and blurry scans.

Finally, each approach comes at costs: tools with an advanced GUI are very flexible to fit different images and adjust what is to be recognized. This, however, crosses with recognizing many images at once in a batch-mode. Batch-mode, on the other hand, increases the risk of erroneous OCR if not all images are of same quality. In sum, the "best tool" for OCR heavily depends on the project's task at hand and the given source material.

For this reason, we will focus our report on comparing tools based on the examples provided by the projects during the workshop by the participants. We will highlight the pros and cons of each tool providing examples of text/table recognition for the files provided.

We have selected the following tools for analysis: Google Docs OCR, Tesseract, ABBYY FineReader, PDF OCR X, Adobe Acrobat. To validate the accuracy, reliability, and performance of the chosen packages several experiments on real data collected from the projects were performed. In Chapter 4 we discuss the experimental setup and the obtained dataset. In Chapter 5 the qualitative OCR results achieved by the OCR packages/services are reported. Subsequently, in Chapter 6 we discuss the results. Chapter 7 contains recommendations based on the outcomes of the analysis. Finally, in Chapter 8 we provide useful links to all the mentioned packages and software tools. The report, thus, gives sufficient information, so that each project may look for the best solution based on their needs.

2. Tools overview

Performance of OCR software packages depend on many factors e.g. language, script, imaging conditions, camera orientation, lighting, quality of the printing, handwritten versus types, cursive versus non-cursive. Also, one must take into account the main goals of the

OCR, whether it is pure OCR (converting documents/images with typed words into machine-readable text) or extracting structured data (layout or tables).

Different OCR engines have different strengths; in this report we therefore explore the capabilities of the following solutions: Google Docs OCR, Tesseract, ABBYY FineReader, PDF OCR X, and Adobe Acrobat. They include free and commercial products and represent high quality solutions in different tasks.

Table 1. Considered OCR packages overview

	OS System	License	Languages	Output Formats	Table recognition	Notes
Tesseract[1]	Windows Mac Linux	Apache	100+	Text, hOCR, PDF, others	No	
Google Drive OCR	Browser	Free	200+	Text	No	Files have to be less than 2 Mb. Output as Google Document.
ABBYY FineReader[2]	Windows Mac Linux	Proprietary	192, +Fraktur	DOC, DOCX, XLS, XLSX, PPTX, RTF, PDF, HTML, CSV, TXT, ODT, DjVu, EPUB, FB2	Yes	
Adobe Acrobat	Windows Mac	Proprietary	46	TXT, PDF, hOCR, XLS, PPTX, XML	Yes	
gImageReader	Windows Mac Linux	GPL	100+	TXT, hOCR, PDF	No	Front End to Tesseract OCR engine
PDF OCR X Community Edition	Windows Mac	Proprietary	100+ + Fraktur	TXT,PDF	No	Drag and drop UI. Works only for one-pagers.
OCRopus	Mac Linux	Apache	All languages using Latin script, + Fraktur	TXT, hOCR, PDF	No	Pluggable framework under active development
Tabula	Windows Mac Linux	MIT	-	CSV	Yes	Requires Java Converts tables to csv only for previously OCRed files
pdfsandwich[3]	Mac Linux	GPL	100+	TXT, PDF	No	Wrapper over tesseract, unpaper

Tesseract is an optical character recognition engine for various operating systems. It is free software, released under the Apache License, Version 2.0, and development has been sponsored by Google since 2006. Tesseract is considered one of the most accurate open-source OCR engines.

Version 4 adds a recurrent neural networks-based OCR engine and models for many additional languages and scripts, bringing the total to 116 languages. Additionally, scripts for 37 languages are supported making it possible to recognize a language by using the script it is written in.

1 https://opensource.google.com/projects/tesseract
2 https://www.abbyy.com/en-eu/finereader/
3 http://www.tobias-elze.de/pdfsandwich/

Tesseract is executed from the command-line interface. While Tesseract is not supplied with a GUI, there are many separate projects which provide a GUI for it. Common examples are OCRFeeder or gImageReader. In addition, R and Python wrapper libraries exist.

Google Docs OCR is an easyto-use and highly available OCR service offered by Google within the Google Drive service. One can convert different types of image data into editable text data using Google Drive. Once we upload an image or a PDF file to the Google Drive, we can start the OCR conversion by right-clicking on the file to select "Open with Google Docs" item, then the image is inside a Google Doc document and the extracted text is right below the image.

One can convert .JPEG, .PNG, .GIF, or PDF (multipage documents) files, though the file size should be 2 MB or less. For best results the text should be at least 10 pixels high. Documents must be right-side up. If the image is facing the wrong way, rotating it before uploading it to Google Drive is necessary.

Google Drive automatically detects the language of the document.

ABBYY FineReader as an advanced OCR software system considered to be the most diverse and functional commercial option available. It has been improving the main functionalities of optical character recognition for many years, providing promising results in text retrieval from digital images. The underlying algorithms of ABBYY FineReader have not yet been illustrated to the research community, and the package is not available as open-source code. Researchers and developers can access the ABBYY FineReader OCR by two different ways:

- The ABBYY FineReader SDK which is available at https://www.abbyy.com/resp/promo/ocr-sdk/, and
- employing a web browser to try it over the Internet at https://finereaderonline.com/en-us/Tasks/Create

Adobe Acrobat Pro is a software suite to foremost work on PDFs but including an OCR system as well. It is used to convert scanned documents, PDF documents, and image documents into editable/searchable documents. It comes in different versions (the most recent version is Acrobat XI Pro). Though it has fewer language options than ABBYY FineReader, Adobe Acrobat Pro is a more pervasive software, partially because it is less academic, and more business-oriented. In addition, on many computers in the CRC it is pre-installed as part of the university's "Grundaustattung".

gImageReader is a simple front-end to *Tesseract*.

On top of the features provided by Tesseract gImageReader includes:

- manual recognition area definition;
- recognized text displayed directly next to the image;
- post-processing the recognized text including spellchecking.

PDF OCR X Community Edition is a simple drag-and-drop utility that converts single-page PDFs and images into text documents or searchable PDF files. It uses advanced OCR technology to extract the text of the PDF even if that text is contained in an image. The Community Edition supports single-page PDFs and images. For multi-page PDFs and batch conversion features one needs to upgrade to the Enterprise Edition.

OCRopus is a free document analysis and optical character recognition system released under the Apache License v2.0 with a very modular design using command-line interfaces. The main components of OCRopus are:

- analysis of the document layout;

» optical character recognition;
» use of statistical language models.

By default, OCRopus comes with a model for English texts and a model for text in Fraktur. These models refer to the script and are largely independent of the actual language. New characters or language variants can be trained either new or in addition. Recent text recognition is based on recurrent neural networks (LSTM) and does not require a language model.

Tabula is a drop and drop tool which allows to extract data in CSV-format, through a simple web interface. Windows & Linux users will need a copy of Java installed. Though, Tabula has some limitations:

» Scanned PDFs: Tabula only works on text-based PDFs, *not scanned documents*.
» Multi-line rows: PDFs with multi-line rows (word wrapped text) are often mis-detected, particularly in tables without graphic row separators.
» Automatic table detection is not available. For now, the user has to do a manual rectangular selection around the candidate tables.

pdfsandwich is a command line tool which is useful for OCR scanned books or journals. It is able to recognize the page layout even for multicolumn text. Essentially, pdfsandwich is a wrapper script which calls the following binaries: unpaper, convert, gs, hocr2pdf, and tesseract. It is known to run on Unix systems and has been tested on Linux and MacOS X. It supports parallel processing on multiprocessor systems.

pdfsandwich provides a number of preprocessing procedures to enhance the quality of the scanned pages before text recognition. Most OCR specific preprocessing options are provided via the program unpaper, such as layout optimization, removal of dark edges, and straightening of skewed scans (deskewing). The layout specification, which is switched off by default, allows the options single for each pdf page containing a single scanned page, and double. Note that in certain circumstances rigid preprocessing options can substantially deteriorate results. For instance, if the wrong layout specifications are provided, whole sub-pages might get filtered out, or figures might be considered visual noise and disappear. If the scan quality is good, it is advisable to completely switch off preprocessing. Simple deskewing of a rotated page (without sub-pages) can be obtained by convert, without any involvement of unpaper.

3. Experimental Setup

For the experiments we have used a dataset consisting of files which each project shared with the group during the workshop. The provided examples are available upon request. Since most of the documents contained multiple pages, we have selected up to four representative pages from each document. This was done to simplify the comparison of the output of each tool.

3.1. Dataset

The following table summarizes the data files used for the comparison and gives additional information about characteristics that affect the results of the OCR.

Table 2. Data characteristics

		Incl. Tables	DPI	Quality	Language
A01-1	world-book-1928.pdf	-	100	scan	English(main) + other
A01-2	world-book-1950.pdf	-		scan	English(main) + other
A01-3	world-book-1968.pdf	-	100	scan	English(main) + other
A01-4	world-book-1988.pdf	-	150	scan	English(main) + other
A02-1	ISSA1949.pdf	textual	300	scan	English
A02-2	ISSA1981.pdf	textual	600	scan	English
A02-3	USSSA_Replacement rates pensions 1965-75.pdf	numeric	200	scan	English
A02-4	USSSA_SSC Expenditures_sel. countries_1990.pdf	numeric	300	scan	English
A02-5	USSSA_World Trends SSB 1935-55.pdf	numeric	300	scan	English
A02-6	USSSA_World Trends SSC 1976.pdf	numeric	200	scan	English
A04-1	Grafiken.pdf	numeric	200	scan	German
A04-2	Foto.jpg	numeric		Handheld photo (blur, skew)	English
A04-3	Seiten aus Reichstagspro-tokolle _1882-83 kopie-4.pdf	-	72	scan	German Fraktur
A04-4	The-Logic-2_Prezworski.pdf	-	300	scan	English
A04-5	trendshealthinsur-ance1968_2015.pdf	numeric	300	Text-based pdf	English
A06-1	Auto_Outline of Foreign Social Insurance.pdf	textual	200	scan	English
A06-2	Outline of Foreign Social In-surance_1940.pdf	textual	200	scan	English
A06-3	Pre_Prim_Geneva_1961.pdf	-	200	scan	English

The provided PDF files include noisy images, blurred images, and multilingual text images. As it can be seen, many files contain tables. Parts of them contain only textual information, others contain numeric data. The majority of the examples are in English, with the exception for a few documents in German and Spanish.

The next table gives an overview about the output formats that were chosen. *Plain text* focuses on extracting only the characters, whereas *hOCR* encodes not only the text, but also style properties such as font family, bold or italic, font size, layout information, recognition confidence metrics and other information using Extensible Markup Language (XML) in the form of Hypertext Markup Language (HTML) or XHTML. *Tables* try to reproduce the table layout and the corresponding cell content.

Table 3. Recommended outputs for each data file

		Text	hOCR	Tables
A01-1	world-book-1928.pdf	+	+	-
A01-2	world-book-1950.pdf	+	+	-
A01-3	world-book-1968.pdf	+	+	-
A01-4	world-book-1988.pdf	+	+	-
A02-1	ISSA1949.pdf	+	-	+

		Text	hOCR	Tables
A02-2	ISSA1981.pdf	+	-	+
A02-3	USSSA_Relplacement rates pensions 1965-75.pdf	+	-	+
A02-4	USSSA_SSC Expenditures_sel.countries_1990.pdf	+	-	+
A02-5	USSSA_World Trends SSB 1935-55.pdf	+	-	+
A02-6	USSSA_World Trends SSC 1976.pdf	+	-	+
A04-1	Grafiken.pdf	+	-	+
A04-2	Foto.jpg	-	-	+
A04-3	Seiten aus Reichstagsprotokolle _1882-83 kopie-4.pdf	+	+	-
A04-4	The-Logic-2_Prezworski.pdf	+	-	-
A04-5	trendshealthinsurance1968_2015.pdf	-	-	+
A06-1	Auto_Outline of Foreign Social Insurance.pdf	+	-	+
A06-2	Outline of Foreign Social Insurance_1940.pdf	+	-	+
A06-3	Pre_Prim_Geneva_1961.pdf	+	+	-

3.1. Preprocessing

We did not perform preprocessing of the provided data (such as layout optimization, removal of dark edges, and deskewing of scans), unless the package tools support them (for example ABBYY FineReader automatically preprocesses files).

Note: For those interested in building their own workflow, the unpaper package is an open source post-processing tool for scanned sheets of paper, especially for book pages that have been scanned from previously created photocopies. The main purpose is to make scanned book pages better readable on screen after conversion to PDF. Additionally, unpaper might be useful to enhance the quality of scanned pages before performing optical character recognition (OCR). Simple deskewing may also be performed using the convert package.

3.2. Output files

All files are stored in subfolders named after each project and are available upon request. In each folder, the naming convention of the files is as follows:

- The original files (.pdf) shared by each project are listed “as is”.
- For each file there may be several processed files with the following extensions: .txt (for plain text), .xlsx (for table extraction), and/or .html (for hOCR). If the output for one file consists of multiple files, they were added to a subfolder, and the folder’s name indicates the output type.
- To distinguish which tool produced the output, the following prefixes have been added to the original file names:
  - ABBY – ABBYY FineReader
  - TESS – Tesseract
  - ADOBE – Adobe Acrobat
  - PDF_OCR_X – PDF OCR X Community Edition
  - TABULA – Tabula

4. Comparative Analysis

We further analyzed and compared the accuracy and reliability of the Google Drive OCR, Tesseract, ABBYY FineReader, and PDF OCR X using the dataset described in Table 2.

We have divided the experiments into three groups based on the output format:

1. Plain text (TXT, Searchable PDF)
2. hOCR (HTML, RTF)
3. Table extraction (CSV, XLS, XLSX)

Plain text focuses on extracting only the characters, whereas hOCR encodes not only the text, but also style properties, such as font family, whether the character is bold, italic, font size, layout information, recognition confidence metrics and other information using Extensible Markup Language (XML) in the form of Hypertext Markup Language (HTML) or XHTML. In table extraction we are interested not only the extraction of characters, but also in the extraction of the table structure.

In the table below it is shown which type of output format is supported by each tool.

Table 4. Supported output format by considered packages

	Plain text	hOCR	Tables
Tesseract	X	X	-
ABBYY	X	X	X
Google Drive OCR	X	-	-
Adobe	-	-	X
PDF OCR X CE	X	-	-
Tabula	-	-	X

4.1. Plain text extraction

In this section we will summarize the findings described in the Plain_text_comparison.docx file. The file contains a detailed comparison of the OCR outcomes for each of the selected files in the dataset.

Text extraction. Adobe, Tesseract and PDF OCR X require some post-processing, including deleting unnecessary line splits and merging words with dashes. This is not an issue, since it can be easily fixed with simple scripts in post-processing. On the other hand, Google OCR and ABBYY do not require such post-processing, though in the case of Google, the words which for example contained dashes are not properly merged, the dashes are replaced with spaces (*con-solidation* becomes *con solidation*), making it harder to fix such errors. Such limitations are neglectable if one only needs a searchable PDF as output, not a plain .txt document.

Being the most flexible open-source and free tool, Tesseract has been more closely examined. It was found though, that the accuracy was quite low and the most common mistakes are recognizing short words as digits (for example, *is – 18* or *15, in – 1n, Minister – IVlinister*).

When analyzing the plain text extraction performance, the software tools can be ranked as follows: ABBYY, Google Drive OCR, Adobe, PDF OCR X, Tesseract. For specific tasks/scripts/settings this ranking may slightly change.

Multilingual files support. Another complication is the ability to select the language of the document and dealing with multi-lingual documents.

Multilingual support is provided only by ABBYY and Tesseract. In Adobe and PDF OCR X only one main language is supported. Google Drive OCR automatically selects the language of the document, so it is unclear, whether it supports multilingual documents. Judging by the outcome for Spanish documents, where the words with accents were not properly recognized, we conclude that only one main language is supported. No multilingual support poses a problem for documents where important information is present in several languages.

Structured layout. Documents containing more complex layouts, for example documents containing tables or multicolumn layouts appeared to be problematic for different tools with Google Drive OCR struggling the most. Though Google Drive OCR was able to extract text segments and properly capture a two-column layout, it outputs a document in which each paragraph was repeated for 2-4 times. This seems to be a bug. Other tools have performed better, with subtle differences in the order of the extracted text segments.

Examples A02-1, A02-2, A02-5, A02-6 represent documents which consist of tables containing textual data. Such structure poses a huge problem for OCR tools in general. Adobe, PDF OCR X and Google Drive OCR approach the document line-by-line; thus, the structure of the table is unrecognized and the results are non-informative and useless. On the other hand, Tesseract and ABBYY are able to extract the structure and retrieve the data from segments, unlike the line-by-line approach. Documents containing tables with textual data will be analyzed in more detail in section 4.3.

Historic fonts. Another challenge faced by the OCR tools was historic font recognition, such as Fraktum or Gothic fonts used in old German documents. Only Google Drive OCR was able to recognize the characters at an adequate level. The second best is PDF OCR X which recognized the font, but with many errors. ABBYY has an extension[4] (not included in the version we have used for testing) which supports historic fonts, and is widely used for this purpose. Adobe and Tesseract completely failed the task, probably due to the models not being trained on such characters.

Fractions extraction. Even though fractions are not common for the majority of the files, it might be useful for some of the projects. In our dataset the file "A04-1 -Grafiken.pdf" contains a table with fractions. Looking at the .xlsx output files using ABBYY and Adobe it can be clearly seen that only Adobe was able to capture the fractions properly. Tesseract has also failed to recognize them, and PDF OCR completely ignores tables, so no results are available.

4.2. hOCR extraction

Within the task of hOCR, we would like to analyze the extraction of not only textual information but also style and layout information. Such information might be useful for further processing of the data, for example splitting the document by country using the country name or other keywords. Splitting just by the country name may result in unnecessary splits if the country is mentioned somewhere else in the text. Splitting by font size or style (if the headers are visibly different, i.e. a greater font size or bold) may also reduce or completely eliminate unwanted splits. Another example would be simply saving the existing segmentation which

4 https://www.frakturschrift.com/en:start

is lost when converting to plain text files. The next table gives an overview for the documents selected for hOCR.

For hOCR, we considered ABBYY, Adobe and Tesseract, since only they support hOCR.

Table 5. Selected documents for plain text extraction

		Comments
A01-1	world-book-1928.pdf	-
A01-2	world-book-1950.pdf	-
A01-3	world-book-1968.pdf	Handwriting or stamp on page messes up output
A01-4	world-book-1988.pdf	-
A06-3	Pre_Prim_Geneva_1961.pdf	-

Capturing fonts styles. Even though Adobe is better at capturing the font styles, Tesseract is better at extracting the paragraphs and layout structure, and quality of the texts per se is better. Yet, ABBYY performs best in terms of both capturing font styles and text quality.

Capturing layout segments ordering. Sometimes the ordering of the text segments is mixed up, for example a table will be added lastly to the page, not in the order as it was in the original text. In Adobe and Tesseract the user cannot influence the ordering, but in ABBYY the user can assign the order of each text fragment for every page of the input file using the user interface.

4.3. Table extraction

The difference in the quality of the text extracted by the OCR tools has been discussed in section 4.1. Here, we will mainly focus on the ability of the considered tools to capture the structure of tables. We considered ABBYY and Adobe, since only they support the extraction of tables. Each tool supports the following outputs:

- the whole file as a single xls file;
- each page as a separate xls file;
- each table as a separate xls file.

To minimize the number of generated files we have selected only the first two available options.

In Table 6. it is shown whether the tool managed to find the tabular structure and properly extract it. It can be seen that in general ABBBYY is able to extract the tables in all the considered examples, even in challenging files such as "A04-2 Foto.jpg", where the file itself represents a handheld photo of a skewed page with blurry edges.

Another important point to be highlighted is that when using Adobe, the user cannot tune the output. Thus, if the software fails to capture the structure properly nothing can be done. This limitation is overcome in ABBYY. Even though ABBYY can detect rows automatically by relying on black separators and white gaps, the interface also allows the user to edit the extracted structure and specify how the table should be divided into rows, for example by adding new separators or merging existing cells. The flip side of this flexibility is the manual adjustment it requires for every table. Depending on the number of tables to be extracted it then becomes a question of weighing the available resources against the efforts the OCR requires.

A final tool we mentioned in the overview is Tabula. Tabula only works on text-based PDFs, *not scanned documents*. Thus, it would work for examples like A04-5, but not for the rest of the examples unless they have already been OCRed. The quality of the extracted

structures for such documents is limited though and comparable to the output produced by Adobe. The next table summarizes our result whereby "+" means that the tool has achieved meaningful results for the document, while "-" means that the outcome was of poor quality.

Table 6. Performance of selected tools for table extraction by file

		ABBYY	ADOBE
A02-1	ISSA1949.pdf	+	-
A02-2	ISSA1981.pdf	+	-
A02-3	USSSA_Relplacement rates pensions 1965-75.pdf	+	+
A02-4	USSSA_SSC Expenditures_sel.countries_1990.pdf	+	+
A02-5	USSSA_World Trends SSB 1935-55.pdf	+	+
A02-6	USSSA_World Trends SSC 1976.pdf	+	-
A04-1	Grafiken.pdf	+	+
A04-2	Foto.jpg	+	-
A04-5	trendshealthinsurance1968_2015.pdf	+	+
A06-1	Auto_Outline of Foreign Social Insurance.pdf	+	-
A06-2	Outline of Foreign Social Insurance_1940.pdf	+	-

5. Conclusion

We performed a qualitative comparative analysis of four OCR solutions, including Google Docs OCR, Tesseract, ABBYY FineReader, and PDF OCR X using a dataset containing typical cases of interest of the workshop participants.

Based on our experimental evaluations using the mentioned dataset without employing advanced image processing procedures (e.g. denoising, image registration), the Google Docs OCR and ABBYY FineReader produced the most promising results in plain text OCR. Though one of the main limitations of Google Drive OCR is the 2Mb file size limit.

In the task of hOCR, ABBYY is considered the best solution in terms of text and layout quality, and it is more sensitive to the font styles than Adobe or Tesseract. The second-best solution is Adobe Acrobat: though the text quality is lower, the extracted layout and styles are close to the results produced by ABBYY. Neither Adobe Acrobat nor Tesseract have the option to set the order of the extracted text fragments of pages manually.

Looking at the results of table extraction, it can be seen that the accuracy of ABBYY is higher than that of Adobe. Most importantly, ABBYY supports a user-interface which allows to easily correct the automatically extracted table structure before outputting the result. This allows more flexibility when dealing with challenging files, though at the cost of preventing batch-processing a large number of tables at once.

In sum, the performance of OCR solutions depends on many factors ranging from the quality of the source material to the objective of OCR. Thus said, there is no "one fits all" solution. Even the most advanced software packages may not be able to perform well in different tasks and settings.

Looking at the results it can be seen that the ABBYY FineReader either outperforms other solutions or provides more flexibility to tune the results to the user's needs. Though it is quite advanced it is one of the most expensive software solutions available. Its quality is underlined by the fact that it is also used by the Staats- und Universitätsbibliothek Bremen. Thus, in case

projects decide to choose ABBYY for their work, they may contact and discuss their task with the library.

If the quality provided by Adobe Acrobat is sufficient, the CRC projects should be able to use it as CRC members should already have a copy installed on their working machines.

When comparing commercial products to open-source solutions it can be seen that there is still a gap in text recognition and flexibility. If the projects decide to proceed with Tesseract, A01 may help in establishing workflows based on short scripts for R or Python and guidelines to simplify the OCR processing and preprocessing of images or recommend already existing user interfaces.

6. Preprocessing Considerations

As we have seen in the experiments, the quality of input images has a great impact on the OCR outputs. All of the examined OCR tools have performed worse when dealing with skewed, blurred, noisy images and texts with fonts smaller than 10pt. Sharp images with even lighting and clear contrasts work best. A resolution of 300 DPI is also recommended. Hence, consider the two most important factors affecting the accuracy of OCR:

» Textual Considerations
Special fonts (typewriter, fraktum), super small fonts (6pt), and low contrast text can all decrease the accuracy of the OCR software. Sometimes, OCR software will not be helpful to use at all.

» Scanning Considerations
Getting a quality image is the first step in having the best and most accurate OCR experience. Consider such things as resolution, brightness, straightness, and discoloration before you scan your text. Rather refrain from performing OCR on handheld photos and low quality images.

Remember that software packages may boast between 97% and 99% accuracy. However, these rates are based on character errors, not word errors!

7. Other Tools and Links

Europeana Newspapers - http://www.europeana-newspapers.eu/public-materials/tools/ - The Europeana Newspapers project has developed a number of free and open source software tools, which help digitize historical newspapers. Projects interested in the OCR of old documents and manual document segmentation should take a look at the tools and interfaces they offer.

Online OCR - https://ocr.space/ - The OCR.space Online OCR service converts scans or (smartphone) images of text documents into editable ones. They claim that the recognition quality is comparable to commercial OCR SDK software. You can test some files online for free, though file size limitations may apply. The engine supports creating searchable PDF, plain text files and json files. Via testing we found out that it struggles with files that have multiple columns and contain tables at the same time. In this case the tool analyses the file line by line, creating useless content.

File A06 - Pre_Prim_Geneva_1961	Adobe	TESS	PDF OCR X	ABBYY	Google Drive OCR
This file example contains only text. The pages are badly cropped, thus the last line is sometimes chopped in half or missing. It can be seen that TESS and PDF OCR X require some post-processing, including deleting unnecessary line splits and merging words with dashes. On the other Adobe, Google OCR and ABBYY do not require such post-processing.	In the other 17 countries there are either only public pre-primary estab-lishments (in 10 countries, namely Albania, Bulgaria, Byelorussia, Czecho-slovakia, Hungary, Poland, Rumania, Ukraine, USSR and Yugoslavia) or only private establishments (in 7 countries, namely Ceylon, Iceland, Lebanon, Federation of Malaya, Nicaragua, Turkey, Union of Burma). Incidentally, in Australia and Canada almost all, and in New Zealand all the establish-ments for children under 5 years of age belong to the private category. As regards the different kinds of establishment, it is more difficult to classify the countries since most of them have different types of institution varying in name but which are much the same in all the countries. If only educational establishments be considered (that is, excluding creches, day nurseries and the homes for children), the most usual terms employed are kindergartens and nursery schools. In a few cases the terms infant schools or infant classes are adopted. Of the 52 countries which speak of kindergartens in their replies, two-thirds apply this term to all their pre-primary establishments. The other r.011ntriPc: 11 11 1lv m lrP rlic::.tlnrtiAn hPtu,PPn thP lrinrlPro!lrtPn - :lnrl thP n11r Pr"	In the other 17 countries there are either only public pre-primary estab-lishments (in 10 countries, namely Albania, Bulgaria, Byelorussia, Czecho-slovakia, Hungary, Poland, Rumania, Ukraine, USSR and Yugoslavia) or only private establishments (in 7 countries, namely Ceylon, Iceland, Lebanon, Federation of Malaya, Nicaragua, Turkey, Union of Burma). Incidentally, in Australia and Canada almost all, and in New Zealand all the establish-ments for children under 5 years of age belong to the private category. As regards the different kinds of establishment, it is more diflicult to classify the countries since most of them have different types of institution varying in name but which are much the same in all the countries. If only' educational establishments be considered (that is, excluding créches, day nurseries and the homes for children), the most usual terms employed are kindergartens and nursery schools. In a few cases the terms infant schools or infant classes are adopted. Of the 52 countries which speak of kindergartens in their replies, two-thirds apply this term to all their pre-primary establishments. The other (‘nnnh‘iPc nmmllv main: 2 dictinr‘finn lwhvr-mn H19 lzinrlprnarfpn and flu: nnrcm-v	In the other 17 countries there are either only public pre-primary estab-lishments (in 10 countries, namely Albania, Bulgaria, Byelorussia, Czecho- slovakia, Hungary, Poland, Rumania, Ukraine, USSR and Yugoslavia) or only private establishments (in 7 countries, namely Ceylon, Iceland, Lebanon, Federation of Malaya, Nicaragua, Turkey, Union of Burma). Incidentally, in Australia and Canada almost all, and in New Zealand all the establish-ments for children under 5 years of age belong to the private category. As regards the different kinds of establishment, it is more difficult to classify the countries since most of them have different types of institution varying in name but which are much the same in all the countries. If only educational establishments be considered (that is, excluding créches, day nurseries and the homes for children), the most usual terms employed are kindergartens and nursery schools. In a few cases the terms infant schools or infant classes are adopted. Of the 52 countries which speak of kindergartens in their replies, two-thirds apply this term to all their pre-primary establishments. The other countries nenally make a dictinetinn hetween the Linderoarten and the nurcerw	In the other 17 countries there are either only public pre-primary establishments (in 10 countries, namely Albania, Bulgaria, Byelorussia, Czechoslovakia, Hungary, Poland, Rumania, Ukraine, USSR and Yugoslavia) or only private establishments (in 7 countries, namely Ceylon, Iceland, Lebanon, Federation of Malaya, Nicaragua, Turkey, Union of Burma). Incidentally, in Australia and Canada almost all, and in New Zealand all the establishments for children under 5 years of age belong to the private category. As regards the different kinds of establishment, it is more difficult to classify the countries since most of them have different types of institution varying in name but which are much the same in all the countries. If only educational establishments be considered (that is, excluding creches, day nurseries and the homes for children), the most usual terms employed are kindergartens and nursery schools. In a few cases the terms infant schools or infant classes are adopted. Of the 52 countries which speak of kindergartens in their replies, two-thirds apply this term to all their pre-primary establishments. The other mnntripc lisnallv mat? n rlictinr-tinn V\pt\x7f»F»r> tin* Vinrlf»rr»art£>n or>A tlif> tinrcpm	In the other 17 countries there are either only public pre-primary estab lishments (in 10 countries, namely Albania, Bulgaria, Byelorussia, Czecho slovakia, Hungary, Poland, Rumania, Ukraine, USSR and Yugoslavia) or only private establishments (in 7 countries, namely Ceylon, Iceland, Lebanon, Federation of Malaya, Nicaragua, Turkey, Union of Burma). Incidentally, in Australia and Canada almost all, and in New Zealand all the establish ments for children under 5 years of age belong to the private category. As regards the different kinds of establishment, it is more difficult to classify the countries since most of them have different types of institution varying in name but which are much the same in all the countries. If only educational establishments be considered that is, excluding crèches, day nurseries and the homes for children), the most usual terms employed are kindergartens and nursery schools. In a few cases the terms infant schools or infant classes are adopted. Of the 52 countries which speak of kindergartens in their replies, two thirds apply this term to all their pre-primary establishments. The other countries usually make a distinction between the kindergarten and the nursery

File A02 -	Adobe	TESS	PDF OCR X	ABBYY	Google Drive OCR
This represents a table containing textual data, thus poses a huge problem to OCR tools. It can be seen that the Adobe, PDF OCR X and Google approach the document line by line, thus the structure of the table is unrecognized and the results are non-informative and useless. On the other hand, Tesseract and ABBYY are able to extract the structure and retrieve the data from cell, unlike the line by line approach. Though the order of the cells may not be correct, though in the case of ABBYY this can be manually adjusted, if necessary, as well as adjusting the table structure(i.e. adding/removing column/row splitters, merging/splitting cells) It can be seen that TESS and PDF OCR X require some post-processing, including deleting unnecessary line splits and merging words with dashes. On the other Adobe, Google OCR and ABBYY do not require such post-processing.	Benefit Amount Qualifying conditions Administration Duration Partial unemployment Qualifying period Waiting period Disqualifications Benefits for total unem- None specified ___ ·-···-··· None required for worker None. No benefit is paid Temporary: Refusal of Minister of Labor and ployment are paid on a currently covered in for 1 day's unemploy- suitable w o r k, 4-13 Social Welfare, general daily basis. general social security ment in week, unless it weeks (longer if re- supervision. National Only full days• of unem- program. is the first or last day peated); voluntary quit- Social Security Office, ployment are normally Persons not currently con- of a 3-day period of un- ting without good rea- public-la w agency in considered days of un- tributing are entitled to employment, eon, 4-13 weeks (longer Ministry, collects conemployment, but half- benefit if able to work if repeated); in com- tributions for all prodays can be added to and if they were mem- plete, in corre ct, or grams and disburses form full days if part of hers of an unemploy- late declaration by in• them to national ada fixed pattern of un- ment fund between Jan- sured, 1-1 3 weeks ministrative bodies. employment due "to in- uary 1, 1938, and May (longer if repeated); ob- Fund for Maintenance termittent work or rota- 10, 1940, or who paid at taining false stamping of the Unemployed, tion of work|	Benefits for total unem—ployrnent are paid on a daily basis. ‘ Only full days/of unem-ployment are normally considered days of un-employment, but half- tion of work; For port—workers, half—days may be considered for benefit. When partial \memploy-ment of 1 day a week occurs, this is considered a day of unemployment when accompanied by at least 2 additional days in same week. Special regulations overn seasonal and, ome workers. The latter must have ‘1 full week of _ unemployment obtain benefit. ‘	Benefit Amount Qualifying conditions Administration oo Duration tent Partial unemployment Qualifying period Waiting period Disqualifications BS as Benefits for totel unem-|None specified.__._..... None required for worker|None. No benefit is paid|Temporary: Refusal of|Minister of Labor -and r Inis- ployment are paid on a currently covered in| for 1 day’s unemploy-| suitable work, 4-13] Social Welfare, general ais daily basis: : general social security] ment in week, unless it} weeks “(longer if . re-| supervision. National “al y| Only -full -days:of unem-| program. is the first or last day| peated); voluntary quit-| Social Security Office, P Ow-| ployment are normally, Persons not currently con-| of a 3-day period of un-|. ting without good rea-| public-law agency in - considered :days_ of un- . tributing are entitled to] employment. son, 4-13 weeks: (longer inistry, collects con- it: 2 employment, .-but®-haif-| benefit if able to work’ if repeated);.incom-| tributions for ‘all pro- days“can -be added to and ‘if they were memi- plete, inoorrect, or). grams © and « disburses auary form fu days if Part of] bers of an unemploy- ate declaration by in-them to national ad-		nsurance, January 1949-Continued Benefit Amount Qualifying conditions Administration Duration Partial unemployment Qualifying period Waiting period Disqualifications Benefits for total unem- None specified ___ ·-···-··· None required for worker None. No benefit is paid Temporary: Refusal of Minister of Labor and ployment are paid on a currently covered in for 1 day's unemploy- suitable w or k, 4-13 Social Welfare, general daily basis. general social security ment in week, unless it weeks (longer if re-supervision. National Only full days• of unem- program. is the first or last day peated); voluntary quit- Social Security Office, ployment are normally Persons not currently con- of a 3-day period of un- ting without good rea- public-law agency in considered days of un- tributing are entitled to employment, eon, 4-13 weeks (longer Ministry, collects con-employment, but half- benefit if able to work if repeated); in com-tributions for all pro- days can be added to and if they were mem-plete, in corre ct, o r grams and disburses form full days if part of hers of an unemploy- late declaration by in• them to national ad- a fixed pattern of un- ment fund between Jan- s u r e d , 1-1 3 w e e k s ministrative bodies. employment due "to in- uary 1, 1938, and May (longer if repeated); ob- Fund for Maintenance termittent work or rota- 10, 1940, or who paid at taining false stamping of the Un employed , tion of work| For port- least 3 months of con- of control card, etc., under Minister, main- workers, half-days may tributions for old-age 1-13 weeks (longer if tains public employ-be considered for benefit. pensions between Jan- repeated); begging, 13 ment offices, conducts

File A04 - Grafiken	Adobe	TESS	PDF OCR X	ABBYY	Google Drive OCR
PDF OCR X ignored the table on the page It can be seen that TESS and PDF OCR X require some post-processing, including deleting unnecessary line splits and merging words with dashes. On the other Adobe, Google OCR and ABBYY do not require such post-processing.	Dass der Bundeskanzler zum Beginn der Legislaturperiode eine Regierungserklärung vor dem Deutschen Bundestag abgibt, ist eine selbstverständliche parlamentarische Tradition. Selbstverständliche Praxis ist überdies geworden, dass die Regierungserklä- rung stets den letzten Akt der Regierungsbildung darstellt. Demnac erfolgt zunächst die Wahl Ernennung und Vereidigung des Bundeskanzlers, dann die Ernennung und V ereidig ng der Bundesminister, und erst im Anschluss daran gibt der Bundeskanzler seine erste Regierungserklärung ab. Bisher haben sich jedenfalls alle Bundeskanzler an diese zeitliche Abfolge gehalten. Keine allgemeine Praxis hat sich allerdings bezüglich der Frage herausgebildet,. innerhalb welcher Frist der Kanzler seine Regierungserklärung abgeben soll. Konrad Adenauer informierte den Bundestag im Jahr 1949 nur fünf Tage nach seiner Wahl über sein Regierungsprogramm, während sich Helmut Kohl 1983 nach seiner Wahl ganze 37 Tage Zeit ließ.	Dass der Bundeskanzler zum Beginn der Legislaturperiode eine Regierungserkl'a'rung vor dem Deutschen Bundestag abgibt, ist eine selbstverständliche parlamentarische Tradition. Selbstverständliche Praxis ist überdies geworden, dass die Regierungserklä- rung stets den letzten Akt der Regierungsbildung darstellt. Demnach erfolgt zunächst die Wahl, Ernennung und Vereidigung des Bundeskanzlers, dann die Ernennung und Vereidigung der Bundesminister, und erst im Anschluss daran gibt der Bundeskanzler seine erste Regierungserklärung ab. Bisher haben sich jedenfalls alle Bundeskanzler an diese zeitliche Abfolge gehalten. Keine allgemeine Praxis hat sich allerdings bezüglich der Frage herausgebildet, innerhalb welcher Frist der Kanzler seine Regierungserklärung abgeben soll. Konrad Adenauer informierte den Bundestag im Jahr 1949 nur fünf Tage nach seiner Wahl über sein Regierungsprogramm, während sich Helmut Kohl 1983 nach seiner Wahl ganze 37 Tage Zeit ließ.	Dass der Bundeskanzler zum Beginn der Legislaturperiode eine Regierungserklärung vor dem Deutschen Bundestag abgibt, ist eine selbstverständliche parlamentarische Tradition. Selbstverständliche Praxis ist überdies geworden, dass die Regierungserklä- rung stets den letzten Akt der Regierungsbildung darstellt. Demnach erfolgt zunächst die Wahl, Ernennung und Vereidigung des Bundeskanzlers, dann die Ernennung und Vereidigung der Bundesminister, und erst im Anschluss daran gibt der Bundeskanzler seine erste Regierungserklärung ab. Bisher haben sich jedenfalls alle Bundeskanzler an diese zeitliche Abfolge gehalten. Keine allgemeine Praxis hat sich allerdings bezüglich der Frage herausgebildet, innerhalb welcher Frist der Kanzler seine Regierungserklärung abgeben soll. Konrad Adenauer informierte den Bundestag im Jahr 1949 nur fünf Tage nach seiner Wahl über sein Regierungsprogramm, während sich Helmut Kohl 1983 nach seiner Wahl ganze 37 Tage Zeit ließ.	Dass der Bundeskanzler zum Beginn der Legislaturperiode eine Regierungserklärung vor dem Deutschen Bundestag abgibt, ist eine selbstverständliche parlamentarische Tradition. Selbstverständliche Praxis ist überdies geworden, dass die Regierungserklärung stets den letzten Akt der Regierungsbildung darstellt. Demnach erfolgt zunächst die Wahl, Ernennung und Vereidigung des Bundeskanzlers, dann die Ernennung und Vereidigung der Bundesminister, und erst im Anschluss daran gibt der Bundeskanzler seine erste Regierungserklärung ab. Bisher haben sich jedenfalls alle Bundeskanzler an diese zeitliche Abfolge gehalten. Keine allgemeine Praxis hat sich allerdings bezüglich der Frage herausgebildet, innerhalb welcher Frist der Kanzler seine Regierungserklärung abgeben soll. Konrad Adenauer informierte den Bundestag im Jahr 1949 nur fünf Tage nach seiner Wahl über sein Regierungsprogramm, während sich Helmut Kohl 1983 nach seiner Wahl ganze 37 Tage Zeit ließ.	Dass der Bundeskanzler zum Beginn der Legislaturperiode eine Regierungserklärung vor dem Deutschen Bundestag abgibt, ist eine selbstverständliche parlamentarische Tradition. Selbstverständliche Praxis ist überdies geworden, dass die Regierungserklä rung stets den letzten Akt der Regierungsbildung darstellt. Demnach erfolgt zunächst die Wahl, Ernennung und Vereidigung des Bundeskanzlers, dann die Ernennung und Vereidigung der Bundesminister, und erst im Anschluss daran gibt der Bundeskanzler seine erste Regierungserklärung ab. Bisher haben sich jedenfalls alle Bundeskanzler an diese zeitliche Abfolge gehalten. Keine allgemeine Praxis hat sich allerdings bezüglich der Frage herausgebildet, innerhalb welcher Frist der Kanzler seine Regierungserklärung abgeben soll. Konrad Adenauer informierte den Bundestag im Jahr 1949 nur fünf Tage nach seiner Wahl über sein Regierungsprogramm, während sich Helmut Kohl 1983 nach seiner Wahl ganze 37 Tage Zeit ließ.

File A04 - Seiten aus Reichstagsprotokolle_1882-83 kopie-4	Adobe	TESS	PDF OCR X	ABBYY	Google Drive OCR
This file proved to be challenging due to the old font used in the document. Google goes line by line but is the only one who recognizes the characters at an adequate level. PDF OCR X also recognized the font and was able to properly segment the page into columns. Adobe, ABBYY and Tesseract completely failed the task. Probably due to the models not being trained on such characters. It can be seen that TESS and PDF OCR X require some post-processing, including deleting unnecessary line splits and merging words with dashes. On the other Adobe, Google OCR and ABBYY do not require such post-processing.	traurige·n2age geraußgi{ft. tt mogitc(}er roeife burc'() bief e @inf djn(tung in ben § 2  udj btt § 1 u1tb bas .Sufta11befommen bes @ef eter5, tu ttt !,lletnen ... m:ugen 1ue1tigftens itt bem § 1 fegt . af5eµtabel . t)t, _gef_agrben fol'tnten. entt bie irage f o ltegt, f o ttJttb f1e 1ebenfaUs burdj unf eren  ef dJfuf3, falle er 1uie er rooUe, nidjt beeinträ  tigt merbett · benn es f djeint fidj ja in ber Stf)at i1n 1uef entl1djen · u1n eit  e Stransµortftage Du f)anbel1t; u1tb bief e 5!::ransµort frage fönne1t mir mit u11f ere1n 9uti_gen . ef fttf3 ttit _itt eilte anbete 2age brinne1t, als b1e1en1ne 1ft, tn bet fte f1dj befinbet uttb bie 1nei1t _ueregrter %reu1tb Dr. SDogrn ge f djilbert gat.	treerigenßage bereuäbiift @ci3iebrttt @ie biefenflteregreebenttber bie @d3[ennnireibe in bee @efet5, unb ee gebt been bereit nieht, fe uerfcbergen @ie eiie übrigen 23erti;eiie, bie wir une nen benz ©efet; errfprecben. fi)aä tft ber preitifdäe @runb, une weici3em iii) heute gegen ben '- 'Jintreg ftimmen werbe: . erfienä weii iii} feine 3eit gehabt habe, mid; ge. infermiren, unb 5meitenä beebeeib, weit ici) glaube, baf3 wir lltfigfiéifiü weife burnb biefe @infcijaitung in ben % 2 audi; ben @ 1 unb base 8uftenbriemmen bee ©efeßrä, wee in meinen biegen wenigftenä in bem @ 1 fein aigentabei ift, gefübrben iii1'ttiten. fitenn bie %?rege fe liegt, fe wirb fie iebenfaiiä bunt; , unferen E3efrbinfi, feiie er wie er weiie, nicht beeintrticijtigt werben; been ee ftijeint fiti) in in ber iibet im wefentiicben'	traurigen Lage heraushilft. Schieben Sie diefen Paragraphen über die ES hlemmfreide in das Gejeg, und es geht dann damit nicht, jo verfcherzen Sie alle übrigen Vortheile, die wir uns von dem Gejet verjprechen. Das ift der praktische Grund, aus welchem ich heute gegen den Antrag flimmen werde: erftens weil ich feine Zeit gehabt habe, mich zu informiren, und zweitens Deshalb, weil ich glaube, daß wir möglicher: weile durch diefe Einjchaltung in den S 2 auch den S 1 umd das AZuftandefommen des Gejeßes, was in meinen Augen wenigftens in dem $ 1 jehr akzeptabel it, gefährden fünnten. Wenn die Frage jo liegt, jo wird fte jedenfalls durch) unseren Beihluß, falle er wie er wolle, nicht beeinträchtigt werden; denn es jcheint fih ja in der That im wejentlichen ' um eine Transportfrage zu handeln; und diefe Transport: frage fünnen wir mit unjerem heutigen Bejchluß nicht in eine andere Lage bringen, als diejenige it, in der fte jich befindet und die mein verehrter Freund Dr.	traurigeitSage Ijerauötjilft. Schieben Sie diefen^aragraphenitber bie Sd)letnmfreibe in baö ©efeh, unb eö geht bann bamit nid)t, fo uerfdjergett Sie alle übrigen 33orfljeile, bie mir unö non bern ©efefj ocrfpredjeit. Saö ift ber praftifdje ©runb, auö meinem ich I;eute gegen bett SIntrag ftimnten inerbe: crfteitö raeil idj feine Seit gehabt hohe, midj ju informiren, nnb graeitenö beöljalb, mcil ich glaube, baf} mir inög(icf)cr= meife burdj biefe (Einfchaltung in beit § 2 audj beit § 1 unb baö Suftanbefointnen beö ©cfegeö, raaö in meinen Slugett menigftenö in bem § 1 febjr afgeptabel ift, gefährbeit fönnten. SSenit bie $rage f° lieflt, fo wirb fie jebenfallö burcf) unferen 23cfdjlufj, falle er mie er raollc, nicht beeinträdjtigt rcerben; beim eö fdjeint fidj ja in ber SCtjat im roefentlidjen um eine Sranöportfrage gu fjanbelit; unb biefe Transport« frage fömteit mir mit unferem heutigen SBefdjlufj nidjt in eine anbere Sage bringen, alö biejeitige ift, in ber fie fid; bcfiitbct unb bie mein uerefjrter Sreunb Dr. Söhnt ge= fdjilbert l)at. (Enblidj aber, meine Herren, ift noch ein ©runb, ber midj beftimmt, gur 3cd gegen ben Eintrag gu ftimmen. Sie fabelt gehört auö bem dftunbe beö Ferrit Sireftorö im Sieidjöfdjajgamte, baff gmifdjen beut beutfdjen Steidje unb ber Sdjmeig citt SOertragönerljältnifj befielt, moitacfj mir bie fdjracigcrifchc Sd) einem foldjen Soll bc; legen, fo ift baö roenigftcuö ber beginn eines 9iüd=	traurigen Lage heraushilft. Schieben Sie diesen Paragraphen über Absaßgebiet zu haben, während unsere Fabrikanten nach die Schlemmkreide in das Gefeß, und es geht dann damit Dänemark und Schweden nicht kommen können und nur ihr nicht, so verscherzen Sie alle übrigen Vortheile, die wir uns eigenes Absaßgebiet haben, wo sie noch die Konkurrenz von von dem Gefeß versprechen. Das ist der praktische Grund, Schweden und Dänemark auszuhalten haben. Das ist es, aus welchem ich heute gegen den Antrag stimmen werde: worüber die Rügenschen Kreidebruchbesißer sich beschweren erstens weil ich keine Zeit gehabt habe, mich zu informiren, und — wie ich meine - fich mit Recht beschweren. Wenn und zweitens deshalb, weil ich glaube, daß wir möglicher in dieser Beziehung etwas geschehen könnte, so würde damit weise durch diese Einschaltung in den § 2 auch den § 1 und der dortigen Industrie ein Dienst geleistet. Ich kann mich das Zustandekommen des Gesekes, was in meinen Augen auf diese Worte beschränken und bemerke, daß der Antrag wenigstens in dem § 1 sehr akzeptabel ist, gefährden fönnten. vorläufig zurückgezogen wird, und wir uns vorbehalten, zur

File A01 - world-book-1988	Adobe	TESS	PDF OCR X	ABBYY	Google Drive OCR
Google couldn't handle this file properly. It was able to extract the textual segments, captured two columns, but for some reason ended up repeating each paragraph 2or 4 times in a row. It can be seen that TESS and PDF OCR X require some post-processing, including deleting unnecessary line splits and merging words with dashes. On the other Adobe, Google OCR and ABBYY do not require such post-processing.	Strategically located between the Middle East, Central Asia, and the Indian subcontinent, Afghanistan is a land marked by physical and social diversity. Physically, the landlocked country ranges from the high mountains of the Hindu Kush in the northeast to low-lying deserts along the western border. Pushtuns (alternatively Pashtuns or Pa-thans) comprise about 55 percent of the population, while Tajiks, who speak Dari (an Afghan variant of Persian), comprise about 30 percent; other groups include Uzbeks, Hazaras, Baluchis, and Turkom.ans. Tribal distinctions (except among the Tajiks) may cut across ethnic cleavages, while religion is a major unifying factor: 90 percent of the people profess Islam (80 percent Sunni and the remainder, mostly Hazara, Shi'ite). In 1980, women constituted approximately 19 percent of the paid work force, with a higher percentage engaged in unpaid agricultural labor. Female participation in gov-ernment is minimal, although there is a women's branch of the ruling party and one female politburo member; among the various mujaheddin resistance groups, female participation is nonexi.stent ..	Strategically located between the Middle East, Central Asia, and the Indian subcontinent, Afghanistan is a land marked by physical and social diversity. Physically, the landlocked country ranges from the high mountains of the Hindu Kush in the northeast to low-lying deserts along the weStern border. PUShIUIIS (alternatively Pashtuns or Pa- thans) comprise about 55 percent of the population, while Ta jiks, who Speak Dari (an Afghan variant of Persian), comprise about 30 percent; other groups include Uzbeks, Hazaras, Baluchis, and Turkomans. Tribal distinctions (except among the Tajiks) may cut across ethnic cleavages, while religion is a major unifying factor: 90 percent of the people profess Islam (30 percent Sunni and the remainder, mostly Hazara, Shi'ite). In 1980, women constituted approximately 19 percent of the paid work force, with a higher percentage engaged in unpaid agricultural labor. Female participation in gov- ernment is minimal, although there is a women's branch of the ruling party and one female politburo member; among the various mujaheddr'n resistance groups, female participation is nonexistent ._	Strategically located between the Middle East, Central Asia, and the Indian subcontinent, Afghanistan is a land marked by physical and social diversity. Physically, the landlocked country ranges from thehigh mountains ofthe HinduKush in the northeast to low-lying deserts along the western border. Pushtuns (alternatively Pashtuns or Pa- thans) compriseabout 55 percent of the population, while Tajiks, who speak Dari (an Afghan variant of Persian), comprise about 30 percent; other groups include Uzbeks, Hazaras, Baluchis, and Turkomans. Tribal distinctions (except among the Tajiks) may cutacross ethmic cleavages, while religion isa major unifying factor: 90 percent ofthe people profess Islam (80 percent Sunniandtheremainder, mostly Hazara, Shi'ite). In 1980, women constituted approximately 19 percent of the paid workforce, with a higher percentage engaged in unpaid agriculturallabor. Female participation in gov- ernment is minimal, although there is awomen's branch of the ruling party and one female politburo member; among the various mujaheddin resistance groups, female participation is nonexistent.	Strategically located between the Middle East, Central Asia, and the Indian subcontinent, Afghanistan is a land marked by physical and social diversity. Physically, the landlocked country ranges from the high mountains of the Hindu Kush in the northeast to low-lying deserts along the western border. Fushtuns (alternatively Pashtuns or Pa-thans) comprise about 55 percent of the population, while Tajiks, who speak Dari (an Afghan variant of Persian), comprise about 30 percent ; other groups include Uzbeks, Hazaras, Baluchis, and Turkomans. Tribal distinctions (except among the Tajiks) may cut across ethnic cleavages, while religion is a major unifying factor: 90 percent of the people profess Islam (80 percent Sunni and the remainder, mostly Hazara, Shiite). In 1980, women constituted approximately 19 percent of the paid workforce, with a higher percentage engaged in unpaid agricultural labor. Female participation in government is minimal, although there is a women's branch of the ruling party and one female politburo member; among the various mujaheddin resistance groups, female participation is nonexistent.	Strategically located between the Middle East, Central Asia, and the Indian subcontinent, Afghanistan is a land marked by physical and social diversity. Physically, the landlocked country ranges from the high mountains ofthe Hindu Kush in the northeast to low-lying deserts along the western border. Pushtuns (alternatively Pashtuns or Pa- thans) comprise about 55 percent of the population, while Tajiks, who speak Dari (an Afghan variant of Persian), comprise about 30 percent; other groups include Uzbeks, Hazaras, Baluchis, and Turkomans. Tribal distinctions (ex.cept among the Tajiks) may cut across ethnic cleavages, while religion is a major unifying factor: 90 percent of the people profess Islam (80 percent Sunni and the rernainder, mostJy Hazara, Shi'ite). In 1980, women constituted approximately 19 percent of the paid work force, with a higher percentage engaged in unpaid agricultural labor. Female participation in gov- ernment is minimal, although there is a women's branch of the ruling party and one female politburo member; among the various mujaheddin resistance groups, female participation is nonexistent..

File A01 - world-book-1968	Adobe	TESS	PDF OCR X	ABBYY	Google Drive OCR
ABBYY did not have Spanish preselected so all special symbols are not recognized. In adobe you can have only one main language for the document. It can be seen that TESS and PDF OCR X require some post-processing, including deleting unnecessary line splits and merging words with dashes. On the other Adobe, Google OCR and ABBYY do not require such post-processing.	President Arturo Frondizi, who assumed office on May 1, 1958, was ousted on March 29, 1962, by a military coup. Although he had been subject to military pressure for some time before that, his overthrow stemmed specifically from the results of the March 18, 1962, elections which constituted a Peronist victory. Frondizi, against military objections, had permitted the Peronists to take part in those elections through candidates of neo-Peronist parties. Following Fron-dizi's overthrow, Jose Maria Guido, Provisional President of the National Senate, was installed as President of the nation under the law of succession, there being no vice-president. A general election was held on July 7, 1963, for the first time in Argentina by proportional representation. Candidates were presented by UDELPA, with P. E. Aramburu as candidate; by the Union Civica Radical Intransigente, led by former President Frondizi (Dr. Oscar Alende was the candidate); the Union Civica Radical del Pueblo, led by Dr. Arturo Illia; and several other smaller parties. The party of Dr. Illia won the election, and his success was hailed as a victory for moderation and a defeat for the forces of former dicta tor Juan D. Peron. The UCRI split before the 1963 elections, half remaining loyal to Alende, and the other half, under Dr. Frondizi, taking the name Movimiento de Inte-gracion y Desarrollo (MID).	President Arturo Frondizi, who assumed oflice on May 1, 1958, was ousted on March 29, 1962, by a military coup. Although he had been subject to military pressure for some time before that, his overthrow stemmed specifically from the results of the March 18, 1962, elections which constituted a Peronist victory. Frondizi, against military objections, had permitted the Peronists to take part in those elections through candidates of neo-Peronist parties. Following Fron-dizi’s overthrow, José Maria Guido, Provisional President of the National Senate, was installed as President of the nation under the law of succession, there being no Vice-president. A general election was held on July 7, 1963, for the first time in Argentina by proportional representation. Candidates were presented by UDELPA, with P. E. Aramburu as candidate; by the Unién Civica Radical Intransigente, led by former President Frondizi (Dr. Oscar Alende was the candidate); the Unién Civica Radical del Pueblo, led by Dr. Arturo Illia; and several other smaller parties. The party of Dr. Illia won the election, and his success was hailed as a Victory for moderation and a defeat for the forces of former dictator Juan D. Perén. The UCRI split before the 1963 elections, half remaining loyal to Alende, and the other half, under Dr. Frondizi, taking the name Movimiento de Inte—gracién y Desarrollo (MID).	Failed	President Arturo Frondizi, who assumed office on May 1, 1958, was ousted on March 29, 1962, by a military coup. Although he had been subject to military pressure for some time before that, his overthrow stemmed specifically from the results of the March 18, 1962, elections which constituted a Peronist victory. Frondizi, against military objections, had permitted the Peronists to take part in those elections through candidates of neo-Peronist parties. Following Fron-dizi’s overthrow, Jose Maria Guido, Provisional President of the National Senate, was installed as President of the nation under the law of succession, there being no vice-president. A general election was held on July 7, 1963, for the first time in Argentina by proportional representation. Candidates were presented by UDELPA, with P. E. Aramburu as candidate; by the Union Civica Radical Intransigente, led by former President Frondizi (Dr. Oscar Alende was the candidate) ; the Union Civica Radical del Pueblo, led by Dr. Arturo Illia; and several other smaller parties. The party of Dr. Illia won the election, and his success was hailed as a victory for moderation and a defeat for the forces of former dictator Juan D. Peron. The UCRI split before the 1963 elections, half remaining loyal to Alende, and the other half, under Dr. Frondizi, taking the name Movimiento de Integration y Desarrollo (MID).	President Arturo Frondizi, who assumed office on May l, 1958, was ousted on March 29, 1962, by a military coup. Although he had been subject to military pressure for some time before that, his overthrow stemmed specifically from the results of the March 18, 1962, elections which constituted a Peronist victory. Frondizi, against military objections, had permitted the Peronists to take part in those elections through candidates of neo-Peronist parties. Following Fron- dizi's overthrow, J ose Maria Guido, Provisional President of the National Senate, was installed as President of the nation under the law of succession, there being no vice-president. A general election was held on July 7, 1963, for the first time in Argentina by proportional representation. Candidates were presented by UDELPA, with P. E. Aramburu as candidate; by the U ni6n Civica Radical Intransigente, led by former President Frondizi (Dr. Oscar Alende was the candidate); the Union Civica Radical del Pueblo, led by Dr. Arturo Illia; and several other smaller parties. The party of Dr. Illia won the election, and his success was hailed as a victory for modera tion and a defeat for the forces of form er dicta tor J uan D. Per6n. The UCRI split before the 1963 elections, half remaining loyal to Alende, and the other half, under Dr. Frondizi, taking the name Movimiento de Inte-graci6n y Desarrollo (MID).

File A02 – USSSA_Relplacement rates pensions 1965-75

Original

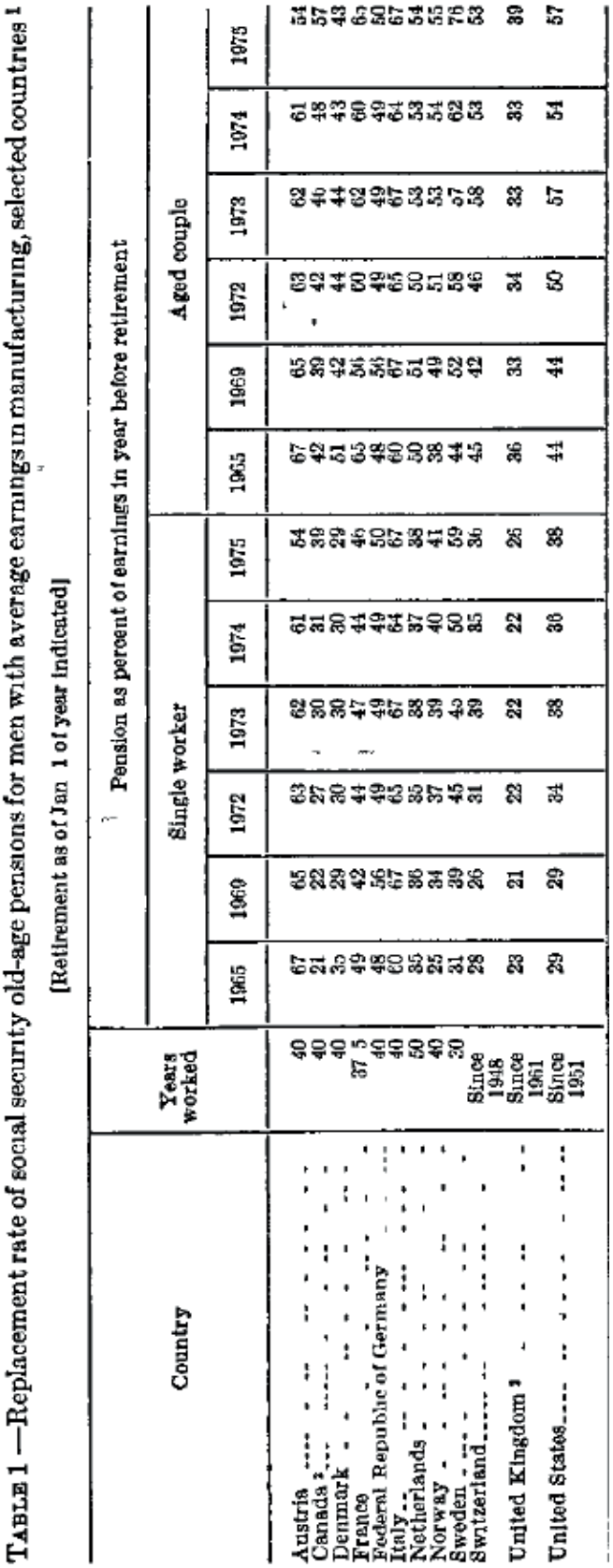

TABLE 1.—Replacement rate of social security old-age pensions for men with average earnings in manufacturing, selected countries [1]

[Retirement as of Jan. 1 of year indicated]

Country	Years worked	Pension as percent of earnings in year before retirement											
		Single worker						Aged couple					
		1965	1969	1972	1973	1974	1975	1965	1969	1972	1973	1974	1975
Austria	40	67	65	63	62	61	54	67	65	63	62	61	54
Canada [2]	40	21	22	27	30	31	39	42	39	42	46	48	57
Denmark	40	35	29	30	30	30	29	51	42	44	44	43	43
France	37.5	49	42	44	47	44	46	65	56	60	62	60	65
Federal Republic of Germany	40	48	56	49	49	49	50	48	56	49	49	49	50
Italy	40	60	67	65	67	64	67	60	67	65	67	64	67
Netherlands	50	35	36	35	38	37	38	50	51	50	53	53	54
Norway	40	25	34	37	39	40	41	38	49	51	53	54	55
Sweden	30	31	39	45	45	50	59	44	52	58	57	62	75
Switzerland	40	28	26	31	39	35	36	45	42	46	58	53	53
United Kingdom [3]	Since 1948	23	21	22	22	22	26	36	33	34	33	33	39
	Since 1961	29	29	34	38	36	38	44	44	50	57	54	57
United States	Since 1951												

Table nested in text. ABBYY outputs the result as a xls file. The file includes the text split line by line as in the original file but the tables are formatted properly if the software was able to recognize them.

Some dots and very small characters were unrecognized, and represented by random symbols. The overall structure of the table is reflected adequately, though.

Table 1 —Replacement rate of social security old-age pensions for men with average earnings m manufacturing, selected countries [1]													
[Retirement as of Jan 1 of year indicated]													
Pension as percent of earnings in year before retirement													
Country	Years worked			Single worker						Aged couple			
		1965	1969	1972	1973	1974	1975	1965	1969	1972	1973	1974	1975
Austria ____ . ... -	40	67	65	63	62	61	54	67	65	~ 03	62	61	54
Canada *...	40	21	22	27	, 30	31	39	42	39	- 42	46	48	57
Denmark.....	40	3o	29	30	30	30	29	51	42	44	44	43	43
France	37 5	49	42	44	* 47	44	46	65	56	60	62	60	6o
Federal Republic of Germany _ _ ...	40	48	56	49	[1] 49	49	50	48	56	49	49	49	50
Italy..	40	CO	67	65	67	64	67	60	67	65	67	64	67
Netherlands.....	50	35	36	35	38	37	38	50	51	50	53	53	54
Norway.....	40	25	34	37	39	40	41	38	49	51	53	54	55
Sweden	£0	31	39	45	4o	50	59	44	52	58	o7	62	76
Switzerland.....	Since 1948	28	26	31	39	35	36	45	42	46	58	53	53
United Kingdom [3].....	Since 1961	23	21	22	22	22	26	36	33	34	33	33	39
United States.....	Since 1951	29	29	34	38	36	33	44	44	50	57	54	57

Original

Table 1.—Expenditures for social security and health care,[1] by country, 1980 and 1983

Country	Share of GNP spent on social security		Percentage change	Social security health component as share of GNP		Share of GNP spent on total health care		Percentage change	Share of GNP spent on social security plus total health care		Percentage change
	(1) 1980	(2) 1983	(3)	(4)[2] 1980	(5)[2] 1983	(6) 1980	(7) 1983 [3]	(10)	(8) 1980	(9) 1983	(10)
Canada	14.8	16.6	12.2	5.3	4.8	7.4	8.7	21.3	16.9	20.5	21.3
United States	12.4	13.6	9.7	4.1	3.0	9.1	10.8	23.0	17.4	21.4	23.0
Netherlands	29.4	33.3	13.3	5.7	6.0	9.1	9.3	11.3	32.8	36.5	11.3
United Kingdom	16.5	19.9	20.6	4.2	4.7	5.8	5.9	16.6	18.1	21.1	16.6
Federal Republic of Germany	24.1	27.4	13.7	5.9	5.9	9.6	9.5	11.5	27.8	31.0	11.5
France	27.5	30.4	10.5	5.5	6.4	8.8	9.2	7.8	30.8	33.2	7.8
Sweden	32.1	34.1	6.2	7.5	8.4	9.4	9.8	4.4	34.0	35.5	4.4
Japan	10.8	12.0	11.1	4.4	4.7	6.0	6.7	12.9	12.4	14.0	12.9

Table nested in text. ABBYY outputs the result as a xls file. The file includes the text split line by line as in the original file but the tables are formatted properly if the software was able to recognize them.

Some dots and very small characters were unrecognized, and represented by random symbols. The overall structure of the table is reflected adequately, though. A row has also been split in two but it does not ruin the overall structure. Such minor issues can be changed in the interface e.g. adding an additional line separator in the table.

Table 1.—Expenditures for social security and health care,' by country, 1980 and 1983

Country	Share of GNP spent on social security		Percentage change	Social security health component as share of GNP		Share of GNP spent on health care		Share of GNP spent on social security plus total health care		Percentage change
	(1) 1980	(2) 1983	(3)	(4)[2] 1980	(5)[2] 1983	(6) 1980	(7) 1983 [3]	(8) 1980	(9) 1983	(10)
Canada.....	14.8	16.6	12.2	5.3	4.8	7.4	8.7	16.9	20.5	21.3
United States.....	12.4	13.6	9.7	4.1	3.0	9.1	10.8	17.4	21.4	23.0
Netherlands.....	29.4	33.3	13.3	5.7	6.0	9.1	9.3	32.8	36.5	11.3
United Kingdom.....	16.5	19.9	20.6	4.2	4.7	5.8	5.9	18.1	21.1	16.6
Federal Republic of Germany.....	24.1	27.4	13.7	5.9	5.9	9.6	9.5	27.8	31.0	11.5
France	27.5	30.4	10.5	5.5	6.4	8.8	9.2	30.8	33.2	7.8
Sweden.....	32.1	34.1	6.2	7.5	8.4	9.4	9.8	34.0	35.5	4.4
Japan.....	10.8	12.0	11.1	4.4	4.7	6.0	6.7	12.4	14.0	12.9

ORIGINAL

***insurance, January 1949*—Continued**

Benefit					Administration
Amount	Duration	Qualifying conditions			
Partial unemployment		Qualifying period	Waiting period	Disqualifications	
Benefits for total unemployment are paid on a daily basis. Only full days of unemployment are normally considered days of unemployment, but half-days can be added to form full days if part of a fixed pattern of unemployment due to intermittent work or rotation of work. For port-workers, half-days may be considered for benefit. When partial unemployment of 1 day a week occurs, this is considered a day of unemployment when accompanied by at least 2 additional days in same week. Special regulations govern seasonal and home workers. The latter must have 1 full week of unemployment to obtain benefit.	None specified	None required for worker currently covered in general social security program. Persons not currently contributing are entitled to benefit if able to work and if they were members of an unemployment fund between January 1, 1938, and May 10, 1940, or who paid at least 3 months of contributions for old-age pensions between January 1, 1935 and May 10, 1940.	None. No benefit is paid for 1 day's unemployment in week, unless it is the first or last day of a 3-day period of unemployment.	Temporary: Refusal of suitable work, 4–13 weeks (longer if repeated); voluntary quitting without good reason, 4–13 weeks (longer if repeated); incomplete, incorrect, or late declaration by insured, 1–13 weeks (longer if repeated); obtaining false stamping of control card, etc., 1–13 weeks (longer if repeated); begging, 13 weeks; notorious misconduct, 13 weeks, (longer if repeated). Absolute: During labor dispute; as result of strike, if strike called with assent of worker.	Minister of Labor and Social Welfare, general supervision. National Social Security Office, public-law agency in Ministry, collects contributions for all programs and disburses them to national administrative bodies. Fund for Maintenance of the Unemployed, under Minister, maintains public employment offices, conducts vocational training, receives funds from National Social Security Office and disburses them to paying agencies. Regional offices of Fund operate in areas named by Minister. The Office and the Fund have equal representation by workers and employers on governing bodies, with a neutral chairman appointed by the Minister. Benefits are paid either by communes or by approved workers' organizations with at least 50,000 members. Appeals are administered by local Claims Commissions, with equal worker and employer representation, and by national Appeals Board named by Minister and having equal worker and employer representation.
No provision in law	12 weeks in 1 year	Not specified in law	8 days	Not specified in law	Ministry of Labor and Social Welfare, supervisory. State Social Insurance Institute, administration. Governing body includes representatives of 3 ministries and of trade unions.

Benefit					Administration
Amount	Duration	Qualifying conditions			
Partial unemployment		Qualifying period	Waiting period	Disqualifications	
Benefits for total unemployment are paid on a daily basis. Only full days of unemployment are normally considered days of unemployment, but halfdays can be added to form full days if part of a fixed pattern of unemployment duetto intermittent work or rotation of work. For port-workers, half-days may be considered for benefit. When paHial unemployment of 11 day a week occurs, this is considered a day of unemployment wheii accompanied by at least 2 additional days in same week. Special regulations govern seasonal and home workers. The latter must have 1 full week of unemployment to obtain benefit.	None specified_____	None required for worker] currently covered in general social security program. Persons not currently contributing are entitled to benefit if able to work and if they were members of an unemployment fund between January 1,1938, and May 10, 1940, or who paid at least 3 months of contributions for old-age pensions between January 1, 1935 and May 10,1940.	'Tone. No benefit is paid for 1 day's unemployment in week, unless it is the first or last day of a 3-day period of unemployment.	Temporary: Refusal of suitable work, 4-13 weeks (longer if repeated) ; voluntary quitting without good reason, 4-13 weeks (longer if repeated); incomplete, inoorrect, or fate declaration by insured, 1-13 weeks (longer if repeated); obtaining false stamping of control card, etc.; 1-13 weeks (longer if repeated); begging, 13 weeks; notorious misconduct, 13 weeks, (longer if repeated). Absolute: During labor dispute; as result of strike, if strike called with assent of worker.	Minister of Labor and Social Welfare, general supervision. National Social Security Office, public-law agency in Ministry, collects contributions for all programs and disburses them to national administrative bodies. Fund for Maintenance of the Unemployed, under Minister, maintains public employment a offices, conducts vocational training, receives funds from National Social Security Office and disburses them to paying agencies. Regional offices of Fund operate in areas named , by Minister. The Office and the . Fund have ,eqUal ; representation by workers and Employers on governing bodies, with a neutral chairman appointed by the Minister. Benefits are paid either by communes or by approved workers* organizations with at least 50,000 members. Appeals are administered by local Claims Commissions, iirith equal worker and: employer representation, and by • national Appeals Board named by Minister and having equal, worker and employer representation.
No provision in law..... **j** =	12 weeks in 1 year **:#W** . : . **lijiiliiiL** \ ■ I" . !..•■ ;,h .!«!;? it:" ?■ 'i-	Not specified in law..... ■ '	**8** days.....	Not specified in law„.....	Ministry of Labor and Social Welfare, superstate Social Insurance Institute, adimnistration. Governing body includes representatives of 3 ministries and of trade unions.

ALBANIA

Cash Benefits for Insured Workers (except permanent disability)	Permanent Disability and Medical Benefits for Insured Workers	Survivor Benefits and Medical Benefits for Dependents	Administrative Organization
Old-age pension: 70% of average earnings during last year or 3 years or highest 3 of last 10 years, inversely according to earnings. Minimum and maximum pensions: 350 and 900 leks a month. Dependents' supplements (if 12 years' continuous work): 20% of pension for 1 dependent. Reduced pension: Proportionate to years of work.	***Invalidity pension:*** 70% of average earnings during last year or 3 years. Minimum and maximum pension, 380 and 700 leks a month. Increment of 10% to 20% for 5–15 years of continuous work for 1 employer; maximum pension, 100% of earnings. Constant-attendance supplement: 15% of earnings. Partial invalidity pension: 40% of earnings, or up to 60% if change in occupation necessary which reduces earnings. Reduced pension if less than full year of coverage.	***Survivor pension:*** 40% of earnings of insured for 1 survivor; 50% for 2 survivors; 65% for 3 or more survivors. Eligible survivors: Spouse and parents if aged, invalid, or caring for orphan; and children, brothers and sisters under age 16 (20 if student, no limit if invalid). Pension divided equally if survivors live apart. Funeral grant: Lump sum of 300 leks.	Ministry of Finance, general supervision. State Social Insurance Office, administration of program through district offices.
Sickness benefit: 70% of earnings, or 85% if 10 years of continuous work for 1 employer. Payable from 1st day of illness for duration of sickness. ***Maternity benefit:*** 75% of earnings, or 95% if 5 years of continuous work for 1 employer. Payable for 5 weeks before and 7 weeks after confinement; may be extended to total of 15 weeks. Birth grant of 280 leks for layette and food.	***Medical benefits:*** Medical services provided directly to patients through facilities of Ministry of Health. Includes medical treatment, accommodation, and medicines in hospital; treatment in home if urgent; office treatment, dental care, and appliances as authorized by regulation.	***Medical benefits for dependents:*** Same medical benefits as insured, except no cost sharing for hospitalization in case of maternity care or for children under age 6. Wife receives same birth grant as working woman.	Ministry of Finance, general supervision. State Social Insurance Office, administration of cash benefits. Ministry of Health, provision of medical benefits with assistance of District Health Services through clinics, hospitals, maternity homes, and other facilities operated by them.
Temporary disability benefit (work injury): 95% of earnings. Payable from 1st day of disability until recovery or certification of permanent disability.	***Permanent disability pension*** (work injury): 80% of average earnings during last year or 3 years, if totally disabled. Constant-attendance supplement: 15% of earnings. Partial disability pension: 50% of earnings. ***Medical benefits*** (work injury): Same as for ordinary sickness above; also appliances.	***Survivor pension*** (work injury): 45% of earnings of insured for 1 survivor; 60% for 2 survivors; 80% for 3 or more survivors. Eligible survivors: Spouse and parents if aged, invalid, or caring for orphan; and children, brothers, and sisters under age 16 (20 if student, no limit if invalid). Pension divided equally if survivors live apart. Funeral grant: Lump sum of 300 leks.	Ministry of Finance, general supervision. State Social Insurance Office, administration of cash benefits. Ministry of Health, provision of medical benefits.

ALBANIA

Cash Benefits for Insured Workers (except permanent disability)	Permanent Disability and Medical Benefits for Insured Workers	Survivor Benefits and Medical Benefits for Dependents	Administrative Organization
Oid^ge pension: 70% of average earnings during last year or 3 yean or highest 3 of last 10 yean, inversely according to earnings. Minimum and maximum pensions: 330 and 900 leks a month. Dependents' supplements Of 12 yean' continuous work): 20% of pension for 1 dependent. Reduced pension: Proportionate to yean of work.	***Invalidity pension:*** 70% of average earnings during last year or 3 years Minimum and maximum pension, 380 and 700 leks a month. Increment of 10% to 20% for 5-15 years of continuous work for 1 employer; maximum pension, 100% of earnings. Constant-attendance supplement: 15% of earnings Partial invalidity pension: 40% of earnings or up to 60% if change in occupation necessary which reduces earnings. Reduced pension if less than full year of coverage.	***Sartivar pension:*** 40% of earnings of insured for 1 survivor; 30% for 2 survivors; 63% for 3 or more survivors Eligible survivors: Spouse and parents if aged, invalid, or caring for orphan; and children, brothers and sisters under age 16 (20 if student, no limit if invalid). Pension divided equally if survivors live apart. Funeral grant Lump sum of 300 leks	Ministry of Finance, general supervision. State Social Insurance Office, administration of program through district offices
Sktness benefit: 70% of earnings, or 83% if 10 yean of continuous work for 1 employer. Payable from 1st day of illness for duration of sickness. **M*smuky kensfit:*** 73% of earnings, or 93% if 3 yean of continuous work for 1 employer. Payable for 3 weeks before and 7 weeks after confinement; may be extended to total of 13 weeks. Birth grant of 280 leks for layette and food.	***Medical benefits:*** Medical services provided directly to patients through facilities of Ministry of Health. Includes medical treatment, accommodation, and medicines in hospital; treatment in home if urgent; office treatment, dental care, and appliances as authorized by regulation.	***Medical benefits for dependents:*** Same medical benefits as insured, except no cost sharing for hospitalization in case of maternity care or for children under age 6. Wife receives same birth grant as working woman.	Ministry of Finance, general supervision. State Social Insurance Office, administration of cash benefits Ministry of Health, provision of medical benefits with assistance of District Health Services through clinics, hospitals maternity homes and other facilities operated by them.
Teoyoearj disokility benefit (work injury): 93% of earnings Payable from 1st day of disability until recovery or certification of permanent disability.	of average earnings during last year or 3 years if totally disabled. Constant-attendance supplement 15% of earnings Partial disability pension: 30% of earnings ***Medical benefits*** (work injury): Same as for ordinary sickness above; also appliances.	***Survivor pension*** (work injury): 45% of earnings of insured for 1 survivor, 60% for 2 survivors; 80% for 3 or more survivors Eligible survivors Spouse and parents if aged, invalid, or caring for orphan; and children, brothers and sisters under age 16 (20 if student, no limit if invalid). Pension divided equally if survivors live apart Funeral grant Lump sum of 300 leks	Ministry of Finance, general supervision. State Social Insurance Office, administration of cash benefits Ministry of Health, provision of medical benefits

ORIGINAL

Bundeskanzler	Datum der Wahl des Bundeskanzlers	Datum der Regierungserklärung	Verzögerung (in Tagen)
Adenauer	15.9.1949	20.0.1949	5
	9.10.1953	20.10.1953	11
	22.10.1957	29.10.1957	7
	7.11.1961	29.11.1961	22
Erhard	16.10.1963	18.10.1963	2
	20.10.1965	10.11.1965	21
Kiesinger	1.12.1966	13.12.1966	12
Brandt	21.10.1969	28.10.1969	7
	14.12.1972	18.1.1973	35
Schmidt	16.5.1974	17.5.1974	1
	15.12.1976	16.12.1976	1
	5.11.1980	24.11.1980	19
Kohl	1.10.1982	13.10.1982	12
	29.3.1983	4.5.1983	37
	11.3.1987	18.3.1987	7
	17.1.1991	30.1.1991	13
	15.11.1994	23.11.1994	8
Schröder	27.10.1998	10.11.1998	14
	22.10.2002	29.10.2002	7

OCR RESULT (ABBYY FINEREADER)

Bundeskanzler	Datum der Wahl des Bundeskanzlers	Datum der Regierungserklärung	Verzögerung (in Tagen)
Adenauer	15.9.1949	20.0.1949	5
	9.10.1953	20.10.1953	11
	22.10.1957	29.10.1957	7
	7.11.1961	29.11.1961	22
Erhard	16.10.1963	18.10.1963	2
	20.10.1965	10.11.1965	21
Kiesinger	1.12.1966	13.12.1966	12
Brandt	21.10.1969	28.10.1969	7
	14.12.1972	18.1.1973	35
Schmidt	16.5.1974	17.5.1974	1
	15.12.1976	16.12.1976	1
	5.11.1980	24.11.1980	19
Kohl	1.10.1982	13.10.1982	12
	29.3.1983	4.5.1983	37
	11.3.1987	18.3.1987	7
	17.1.1991	30.1.1991	13
	15.11.1994	23.11.1994	8
Schröder	27.10.1998	10.11.1998	14
	22.10.2002	29.10.2002	7

Original

Tab. 119, 120

Gesetzliche Krankenversicherung (neue Bundesländer)

Beitragssätze für Pflichtmitglieder mit Entgeltfortzahlungsanspruch[1]

1. Januar 1993

Beitragssatz in % des Grundlohnes	Krankenkassen insgesamt[2]		Ortskrankenkassen		Betriebskrankenkassen		Innungskrankenkassen		See-Krankenkasse		Bundesknappschaft		Ersatzkassen für Arbeiter		Ersatzkassen für Angestellte	
	Kassen	Mitglieder	Kassen	Mitglieder	Kassen	Mitglieder	Kassen	Mitglieder	Kassen	Mitglieder	Kassen	Mitglieder	Kassen	Mitglieder	Kassen	Mitglieder
9,2	1	68	—	—	1	68	—	—	—	—	—	—	—	—	—	—
9,5	1	857	—	—	1	857	—	—	—	—	—	—	—	—	—	—
9,6	1	100	—	—	1	100	—	—	—	—	—	—	—	—	—	—
9,8	3	258	—	—	3	258	—	—	—	—	—	—	—	—	—	—
10,0	3	974	—	—	3	974	—	—	—	—	—	—	—	—	—	—
10,2	7	7506	—	—	7	7506	—	—	—	—	—	—	—	—	—	—
10,3	1	552	—	—	1	552	—	—	—	—	—	—	—	—	—	—
10,4	5	8806	—	—	5	8806	—	—	—	—	—	—	—	—	—	—
10,5	7	3456	—	—	7	3456	—	—	—	—	—	—	—	—	—	—
10,6	11	11918	—	—	11	11918	—	—	—	—	—	—	—	—	—	—
10,7	6	3312	—	—	6	3312	—	—	—	—	—	—	—	—	—	—
10,8	5	1867	—	—	5	1867	—	—	—	—	—	—	—	—	—	—
10,9	6	2368	—	—	6	2368	—	—	—	—	—	—	—	—	—	—
11,0	16	232024	—	—	15	30808	—	—	—	—	—	—	—	—	—	—
11,1	2	14911	—	—	1	100	—	—	—	—	—	—	—	—	—	—
11,2	6	7160	—	—	6	7160	—	—	—	—	—	—	—	—	1	201216
11,3	2	1269	—	—	2	1269	—	—	—	—	—	—	—	—	1	14811
11,4	7	2929	—	—	7	2929	—	—	—	—	—	—	—	—	—	—
11,5	15	268462	—	—	14	233599	—	—	—	—	—	—	—	—	—	—
11,6	5	27191	—	—	5	27191	—	—	—	—	—	—	1	34863	—	—
11,7	1	7501	—	—	1	7501	—	—	—	—	—	—	—	—	—	—
11,8	24	172501	—	—	17	34729	7	137772	—	—	—	—	—	—	—	—
11,9	15	38286	—	—	12	12421	1	12446	1	8037	—	—	1	5382	—	—
12,0	3	1643	—	—	3	1643	—	—	—	—	—	—	—	—	—	—
12,1	2	27546	—	—	1	25	1	27521	—	—	—	—	—	—	—	—
12,2	8	495509	—	—	4	128060	3	39717	—	—	—	—	—	—	1	327732
12,3	14	2014367	—	—	2	8630	9	112876	—	—	—	—	—	—	3	1892861
12,4	3	20783	—	—	2	137	1	20646	—	—	—	—	—	—	—	—
12,5	1	5341	—	—	—	—	1	5341	—	—	—	—	—	—	—	—
12,6	2	26441	—	—	1	11480	1	14961	—	—	—	—	—	—	—	—
12,7	1	55045	—	—	1	55045	—	—	—	—	—	—	—	—	—	—
12,8	24	3449204	12	3232262	3	875	7	65535	—	—	1	150108	1	424	—	—
12,9	1	648	—	—	1	648	—	—	—	—	—	—	—	—	—	—
13,3	1	2176	—	—	1	2176	—	—	—	—	—	—	—	—	—	—
13,5	1	9	—	—	1	9	—	—	—	—	—	—	—	—	—	—
13,7	2	6931	—	—	—	—	—	—	—	—	—	—	2	6931	—	—
13,8	1	606873	1	606873	—	—	—	—	—	—	—	—	—	—	—	—
Insges.	214	7526792	13	3839135	157	608477	31	436815	1	8037	1	150108	5	47600	6	2436620
						Durchschnittlicher Beitragssatz in % des Grundlohnes[3]										
Insges.		12,54		12,96		11,73		12,20		11,90		12,80		11,88		12,17

1) Für mindestens 6 Wochen. – 2) Ohne landwirtschaftliche Krankenkassen, die den Beitragssatz nicht in % des Grundlohnes festsetzen. – 3) Mit der Mitgliederzahl gewogener Durchschnitt.

Hinweis: Daten von Juli 1992 wurden im Heft 10/92 veröffentlicht.

Quelle: Bundesministerium für Arbeit und Sozialordnung

BMA-Ib4

Gesetzliche Krankenversicherung (neue Bundesländer)

Durchschnittliche Beitragssätze für Pflichtmitglieder

I2£	Krankenkassen insgesamt[2,] Kassen	Mit-glieder	Oits-krankcnkassen Kassen	Mit-glieder	Betriebs-krankenkassen Kassen	Mit-glieder	Inoangs-krankenkassen [Kassen	Mit-; giieder	See-Krankenkasse Kassen	Mit-gliede r	Brmdes-knappschaft | Kassen	Mit-glieder
	1	68			1	68			1-			
ii	1	857	—-	—	t	857			1 «■—	—	—	—
i	1	100	—	—	1	100	_		_	—*	—	—
I 0,5	3	2S8	—	—	3	258		•am		—	—	—
I 1 *ido*	3	974	—		3	974		-		_	—	—
*v' F	7	7506	—	—			———			—	• -*	—
ift2					7	7506		_	_			
Jfti	1	552	—		1	552		_	_	—	—	—
{0,4	5	8806	—	—	5	8806			_	—-	—	—
/0,5	7	3456	—	—	7	3456		_	_	—	—	—
70,*	11	11918	—		11	11918		_	_	—	—	—
70,«	5	1867	—		5	1867	—	_	_	—	—	—
10,9	6	2368	—	—	6	2368	_,	_	_	—	—	—
11.0	16	232024	—		15	30808	—	■	_	_	—	—
11,i	2	14911	—	—-	1	100	_	_	—	—	—	—
11.8	24	172501	—	—	17	34729	7	137772	—	—	—	—
11.9	15	38286	_	—	12	12421	1	12446	i	8037	—	—
12.0	3	1643	—	—*	3	1643	_	_	_	—	—	—
12,1	2	27546	—	—	1	25	1	27521	—	—	—	—
12,2	8	495509	—	—	4	128060	3	39717	_	—	—	—
12,3	14	2014367	—	—*	2	8630	9	112876	—	—	—	—
12,6	2	26441	—	—	1	11480	1	14961	—	—	—	—
12,7	1	55045	—		1	55045	—	—	—	—	—	—
12,8	24	3449204	12	3232262	3	875	7	65535	—	_ —	1	150108
12,9	1	648	—	‘ -	1	648	—	—	—	—	—	—
13,3	1	2176	—	—	1	2176	—	—	—	—	—	—
13,5	1	9	—	—	1	9	—	—	—	—	—	—
Insges.	214	7526792	13	3839135	157	608477	31	436815	i	803		150
										7 1	1	108
Insges.	*12,54*		*12,96*		Durchschn *11,73*		ttlicher Beitragssatz *12,20*					

ORIGINAL

Table 1. Percentages (and standard errors) of persons under 65 years of age with health insurance coverage, by coverage type, and without health insurance: United States, selected years 1968-2015

Year	Sample size	Private coverage (any)[1]	Private coverage (employer)[2]	Private coverage (other)[3]	Medicaid	Medicare	Other public coverage	Uninsured[4]
1968	120,670	79.3 (0.39)	---	---	---	---	---	---
1970	44,373	78.7 (0.53)	68.6 (0.60)	10.0 (0.37)	---	---	---	---
1972	119,939	77.3 (0.39)	69.4 (0.43)	7.8 (0.18)	3.5 (0.14)	---	2.6 (0.18)	16.7 (0.32)
1974	104,727	79.7 (0.31)	70.5 (0.35)	9.6 (0.18)	4.7 (0.16)	---	2.5 (0.20)	13.1 (0.24)
1976	101,594	78.9 (0.31)	68.5 (0.32)	10.3 (0.19)	4.9 (0.16)	0.2 (0.02)	2.6 (0.19)	14.1 (0.24)
1978	98,465	79.3 (0.34)	70.2 (0.35)	9.2 (0.19)	6.7 (0.19)	1.2 (0.04)	2.3 (0.16)	12.0 (0.22)
1980	91,425	79.4 (0.38)	71.4 (0.40)	8.0 (0.20)	7.1 (0.19)	1.4 (0.05)	2.0 (0.16)	12.0 (0.26)
1982	92,489	78.1 (0.53)	70.3 (0.55)	7.9 (0.21)	6.1 (0.29)	1.2 (0.04)	3.7 (0.21)	13.9 (0.36)
1984	46,729	76.9 (0.64)	68.4 (0.67)	8.7 (0.27)	6.8 (0.34)	1.1 (0.06)	3.6 (0.26)	14.6 (0.46)
1986	93,396	76.7 (0.62)	69.1 (0.62)	7.7 (0.21)	6.8 (0.33)	1.2 (0.04)	3.7 (0.23)	14.5 (0.39)

Table 1. Percentages (and standard errors) of persons under 65 years of age with health insurance coverage, by coverage type, and without health insurance: United States, selected years 1968-2015

Year	Sample size	Private coverage (any)[1]	Private coverage (employer)[2]	Private coverage (other)[3]
1968	120,670	79.3 (0.39)	—	—
1970	44,373	78.7 (0.53)	68.6 (0.60)	10.0 (0.37)
1972	119,939	77.3 (0.39)	69.4 (0.43)	7.8 (0.18)
1974	104,727	79.7 (0.31)	70.5 (0.35)	9.6 (0.18)
1976	101,594	78.9 (0.31)	68.5 (0.32)	10.3 (0.19)
1978	98,465	79.3 (0.34)	70.2 (0.35)	9.2 (0.19)
1980	91,425	79.4 (0.38)	71.4 (0.40)	8.0 (0.20)
1982	92,489	78.1 (0.53)	70.3 (0.55)	7.9 (0.21)
1984	46,729	76.9 (0.64)	68.4 (0.67)	8.7 (0.27)
1986	93,396	76.7 (0.62)	69.1 (0.62)	7.7 (0.21)
1989	54,860	76.8 (0.71)	69.3 (0.76)	7.6 (0.33)
1990	102,684	75.9 (0.51)	68.3 (0.51)	7.6 (0.19)
1991	105,053	74.2 (0.43)	66.4 (0.47)	7.8 (0.28)
1992	105,316	73.6 (0.48)	62.8 (0.52)	10.8 (0.31)
1993	113,042	72.0 (0.46)	64.9 (0.45)	7.1 (0.18)
1994	101,608	69.9 (0.50)	64.0 (0.48)	5.9 (0.17)
1995	90,512	71.3 (0.42)	65.6 (0.43)	5.7 (0.16)
1996	56,268	71.2 (0.55)	65.1 (0.57)	6.1 (0.22)
1997	91,275	70.7 (0.36)	66.4 (0.36)	4.2 (0.13)
1998	87,020	72.1 (0.36)	67.5 (0.37)	4.6 (0.14)

ORIGINAL

Chart I.—*Compulsory old-age, invalidity,*

Country and year law enacted	Coverage	Contributions		Government	Old-age pension	
		Insured	Employer		Conditions for receipt	Amount
Australia: 1938 (not yet effective).	Employed persons aged 16–65 in the case of men, 16–60 in the case of women. *Important exceptions.*—Nonmanual employees at a rate of remuneration exceeding £365 a year.	Shared equally by insured and employer. Weekly rate for sickness (chart II) and invalidity 1—men, 1s. 3d.; women, 1s. 2d.		Subsidy for sickness and invalidity insurance—£100,000 a year for administration; plus 10s. per insured per year until deficit of initial generation is paid for.	----------	----------
Belgium: 1924	Wage earners over age 14.	Varying with wage classes and shared equally by insured and employer.		Subsidy for each pension; funds required for original pensions and transitional bonuses.	Age 65; may be claimed by men at age 60 and by women at age 55. No qualifying period.	Basic amount equivalent to contributions. Plus Government subsidy of 50 percent of pension, increased for persons born before 1884 and reduced if pension is claimed before age 65. Maximum Government subsidy 1,400 francs a year. Bonus for transitional generation.
1925	Salaried employees	3 percent of salary up to 18,000 francs a year.	4 percent of salary up to 18,000 francs a year until 1966; increasing thereafter and reaching 5 percent in 1991.	Subsidy for each pension; funds required for transitional bonuses.	Age 65 for men, 60 for women; may be claimed 10 years earlier. No qualifying period.	Basic amount equivalent to contributions. Plus Government subsidy of 50 percent of pension, increased for persons born before 1884 and reduced if pension is claimed before age 65. Maximum Government subsidy 1,200 francs a year. Bonus for transitional generation.

Country and year law enacted	Coverage	Contributions			Old-age pension	
		Insured	Employer	Government	Conditions for receipt	Amount
Australia: 1938 (not yet effective).	Employed persons aged 16-65 in the case of men, 16-60 in the case of women. *Important ex* elusion.—Nonmanual employees at a rate of remuneration exceeding £365 a year.	Shared equally by insured and employer. Weekly rate for sickness (chart II) and invalidity »—men, Is. 3d.; women, Is. 2d.		Subsidy for sickness and invalidity insurance— £ 100, (MX) a ye>ar for administration; plus 10s. per insured per year until deficit of initial generation is paid for.		
Belgium: 1924 ...	Wage earners over age 14.	Varying with wage classes and shared equally by insured and employer.		Subsidy for each pension; funds required for orphans' pensions and transitional bonuses.	Age 65; may be claimed by men at age 60 and by women at age 55. No qualifying period.	Basic amount equivalent to contributions. Plus Government subsidy of 50 percent of pension, increased for persons born before 1884 and reduced if pension is claimed before age 65. Maximum Government subsidy 1,200 francs a year. Bonus for transitional generation.
1925....	Salaried employees ...	3 percent of salary up to 18,000 francs a year.	4 percent of salary up to 18,000 francs a year until 1960; increasing thereafter and reaching 5 percent in 1991.	Subsidy for each pension; funds required for transitional bonuses.	Age 65 for men, 60 for women; may be claimed 10 years earlier. No qualifying period.	Basic amount equivalent to contributions. Plus Government subsidy of 50 percent of pension, increased for persons born before 1884 and reduced if pension is claimed before age 65. Maximum Government subsidy 1,200 francs a year. Bonus for transitional generation.