Die Texterkennung als Herausforderung bei der Digitalisierung von Tabellen

Eine Beschreibung des Projektes für belgische historische Zählungen (KU Leuven Libraries Economics and Business)

André Davids, KU Leuven

Zusammenfassung
Bereits seit mehr als 5000 Jahren finden Volkszählungen statt. Während sie ursprünglich nur zur Steuererhebung und zu militärischen Zwecken durchgeführt wurden, dienten sie später auch der wissenschaftlichen Forschung. Die ersten Zählungen, die von Anfang an auch der Forschung zur Verfügung standen, wurden 1846 unter der Leitung von Adolphe Quetelet in Belgien durchgeführt. Diese Zählungen wurden in regelmäßigen Abständen wiederholt. Da die Analyse dieser Zählungen aufgrund ihres Umfanges und ihres Formats sehr aufwendig ist, ist es sinnvoll, die dort enthaltenen Informationen mit Hilfe von Retrodigitalisierung und elektronischer Texterkennung aufzubereiten. Die wirtschaftswissenschaftliche Bibliothek der KU Leuven (Belgien) arbeitet zur Zeit an einem Projekt, das sich zum Ziel gesetzt hat, die gedruckten Ausgaben der belgischen Industriezählungen von 1846 bis 1947 als Excel-Kalkulationstabellen anzubieten. In diesem Artikel wird sowohl auf die damit verbundenen Herausforderungen eingegangen als auch die Arbeitsweise beschrieben.

Summary
Censuses have been taking place for more than 5000 years. They were originally carried out for the purposes of tax collection and the military, but later also used for scientific research. The first censuses which, from the start, were also available for research were carried out in Belgium in 1846 under the direction of Adolphe Quetelet. After Quetelet's famous first censuses for research, many subsequent censuses were taken. Since the analysis of these censuses is very time-consuming due to their extent and format, it makes sense to convert them into digital form by means of electronic text recognition. The KU Leuven Libraries Economics and Business (Belgium) is currently working on a project aimed at offering the printed editions of the Belgian industrial censuses between 1846 and 1947 as Excel spreadsheets. This article addresses the challenges involved and describes the procedures.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/5584
Autorenidentifikation: Davids, André: ORCID https://orcid.org/0000-0002-3333-8564
Schlagwörter: Digitalisierung; Texterkennung; OCR

1. Einleitung

Seit 2017 wird an der Bibliothek für Wirtschaftswissenschaften der KU Leuven an einem Projekt gearbeitet, mit dem Ziel die veröffentlichten Ausgaben belgischer Zensusdaten zu digitalisieren.

Die Texterkennung von digitalisierten Tabellen bringt im Vergleich mit der Verarbeitung von herkömmlichen Fließtexten zusätzliche Herausforderungen mit sich.

Im vorliegenden Artikel wird die Arbeitsweise des Projekts vorgestellt. Hauptaugenmerk gilt der elektronischen Texterkennung mittels OCR-Software. Alternative Verfahren wie Double Keying und Crowdsourcing werden nicht weiter besprochen.

In Kapitel 2 wird kurz auf die Geschichte der Zählungen eingegangen. In Kapitel 3 werden der potenzielle Nutzen der digitalen Aufbereitung von Volkszählungsdokumenten und die verschiedenen möglichen Bearbeitungsniveaus erläutert. Kapitel 4 beschreibt konkret die Arbeitsweise an der KU Leuven. In Kapitel 5 wird nochmals auf die größten Herausforderungen eingegangen und ein Ausblick auf die kommende Arbeit gegeben.

2. Zählungen in Belgien

Menschen werden bereits seit tausenden Jahren gezählt. Früheste Zeugnisse darüber finden sich in Ägypten, wo bereits ca. 2700 v. Chr. Steuerlisten erstellt wurden. Eine weitere Volkszählung lässt sich auch um ca. 1100 v. Chr. in Ägypten nachweisen.[1] Auch wenn die Historizität nicht nachgewiesen ist, so berichtet das Buch Numeri im Alten Testament der Bibel von zwei Zählungen, die die Israeliten nach dem Auszug aus Ägypten und nach deren vierzigjährigem Aufenthalt in der Wüste durchgeführt haben. Seit dem 6. Jahrhundert v. Chr. führten die Römer regelmäßig Volkszählungen durch.[2] Die bekannteste römische Volkzählung ist diejenige, die um das Jahr 0 durch Kaiser Augustus in Auftrag gegeben wurde, um alle Bewohner des Römischen Reiches in Steuerlisten zu erfassen. Hierüber berichtet die Bibel im Lukasevangelium (Lk 2,1-5). Die ersten modernen Zählungen fanden 1528 in Litauen und ab 1686 in Schweden statt. Nach 1800 wurden in den meisten europäischen Ländern regelmäßige Volkszählungen durchgeführt.

Bis Mitte des 19. Jahrhunderts dienten Volkzählungen fast ausschließlich administrativen Zwecken, wie zum Beispiel der Steuererhebung, der Ermittlung der wehrpflichtigen Männer oder der Festlegung von Sachabgaben im Kriegsfalle.

Die ersten Zählungen auf dem Gebiet des heutigen Belgien wurden bereits vor der Staatsgründung (1830) von Frankreich und dem Königreich der Vereinigten Niederlande durchgeführt.

Frankreich annektierte das belgische Grundgebiet nach der französischen Revolution im Jahr 1794. Diese Zeit zeichnete sich durch eine sehr aktive statistische Tätigkeit aus, die jedoch sehr ungeordnet verlief. Es gab keine zentral organisierten Zählungen in Frankreich. Jedes Département und verschiedene Ministerien führten nach Bedarf eigene Zählungen durch. Eine der bekanntesten Zählungen

1 Černý, Jaroslav: Consanguineous Marriages in Pharaonic Egypt, in: Journal of Egyptian Archeology 40, 1954, S. 28–29. Online: <https://www.jstor.org/stable/3855544>.
2 Tenney, Frank: Roman Census Statistics from 508 to 225 B.C., in: The American Journal of Philology 51 (4), 1930, S. 313–324. Online: <https://www.jstor.org/stable/289892>.

ist die Bevölkerungsliste des Jahres IV (1796), in der zum ersten Mal die Bevölkerung der belgischen Départements aufgeführt wurde.[3]

Von 1815 bis 1830 gehörte das heutige Belgien zum Königreich der Vereinigten Niederlande. In den ersten Jahren wurden auch hier sehr ungeordnet Zählungen durchgeführt, um die „neue" Bevölkerung kennenzulernen, Steuern erheben zu können, usw. Erst 1829/30 wurde eine allgemeine Volkszählung organisiert, die aber wegen der belgischen Revolution (1830) auf dem Grundgebiet des heutigen Belgiens nicht abgeschlossen wurde. Der Statistiker Edouard Smits, der an dieser niederländischen Volkszählung mitgearbeitet hatte, versuchte sie später zusammen mit Adolphe Quetelet und dem belgischen Staat zu rekonstruieren.[4]

Diese französischen und niederländischen Zählungen wurden von vielen belgischen Gemeinden durch die bewusste Angabe falscher Zahlen sabotiert, um Steuern und mögliche kriegswichtige Abgaben (Wehrpflichtige, Pferde, Korn, …) zu minimieren.[5]

Die ersten allumfassenden belgischen Zählungen seit der Staatsgründung waren auch die ersten, die im Gegensatz zu Zählungen in anderen Ländern nicht mehr nur für allgemeine administrative, sondern auch für wissenschaftliche Zwecke bestimmt waren. Sie fanden im Auftrag des belgischen Staates statt und wurden 1846 unter der Leitung des Begründers der wissenschaftlichen Sozialstatistik, Adolphe Quetelet (1796-1874), durchgeführt. Sie bestanden aus drei Teilen: Volkszählung, Landwirtschaftszählung sowie Industrie- und Berufezählung.

Diese Zählung hatte für ganz Europa Modellcharakter. Der niederländische Historiker Paul Klep nennt sie sogar die beste Zählung Europas im 19. Jahrhunderts: „*De Belgische Algemene Telling van 1846 - met op de achtergrond de grote statisticus Adolphe Quetelet - is wat mij betreft de meest uitgebreide en kwalitatief allerbeste van alle Europese tellingen in de 19de eeuw.*" (Die belgische Volkszählung von 1846 – mit dem großen Statistiker Adolphe Quetelet im Hintergrund – ist meiner Meinung nach die umfangreichste und qualitativ hochwertigste aller europäischen Volkszählungen im 19. Jahrhundert.)[6]

Eine zweite bedeutende Zählung dieser Zeit war die Industriezählung vom 31. Oktober 1896, die in Umfang und Detailreichtum weltweit einmalig war. Der französische Statistiker Lucien March schrieb im Jahre 1902: „*…qu'aucune enquête professionnelle générale n'a encore été ni plus instructive, ni plus complète, à un égal degré de précision et d'exactitude.*" (… dass es noch keine professionelle allgemeine Umfrage gab, die informativer und vollständiger ist, und das in diesem Maße an Präzision und Genauigkeit.)[7]

3 Bracke, Nele: Bronnen voor de industriële geschiedens. Gids voor Oost-Vlaanderen (1750-1945), Gent 2000. S. 198.
4 Bracke, Nele: Een monument voor het land. Overheidsstatistiek in België 1795-1870, Gent 2008. Online: <https://www.oapen.org/search?identifier=366390>.
5 Ebd.
6 Klep, Paul: Politieke strubbelingen rond de volkstelling 1859, Voorburg, 2007. Online: <http://www.volkstelling.nl/nl/documentatie/1859/rede_pklep/index.html>.
7 March, Lucien: Le recensement des industries en Belgique en 1896, in: Journal de la société statistique de Paris 43, 1902, S. 265.

Volkszählungen werden in Belgien seit 1846 ungefähr alle zehn Jahre durchgeführt. Bis 1991 wurde jedem Haushalt ein Zählformular zugesandt, welches verpflichtend ausgefüllt werden musste. Diese Formulare wurden von den Gemeinden ausgewertet. Die Gemeindezählungen wurden daraufhin von den Provinzen kontrolliert, worauf der belgische Staat dann die Gesamtzahlen ermittelte.

Die Volkzählung des Jahres 2001 war bereits keine klassische Zählung mehr. Die Bevölkerungsanzahl wurde auf Basis des Nationalregisters ermittelt. Jeder Haushalt erhielt nur noch ein Formular mit sozioökonomischen Fragen. Seit 2011 wird die Volkszählung ausschließlich auf Grundlage verschiedener Verwaltungsregister durchgeführt. Die Bevölkerung wird seitdem nicht mehr persönlich befragt.

3. Warum digitalisieren?

Am Beispiel der Industriezählung des Jahres 1896[8], dokumentiert in 18 Bänden mit insgesamt ungefähr 12.000 Seiten, wird schnell deutlich, dass sich die gedruckten Ausgaben nicht zur schnellen Recherche oder komplexeren Berechnungen eignen. Zu umfangreich, zu detailliert und zu unübersichtlich sind die Tabellen. Obwohl diese Industriezählung als Meisterwerk angeführt wird, wurde kaum Forschung mit den dort dokumentierten Inhalten betrieben.

In der gedruckten Ausgabe können Informationen nachgeschlagen werden, die als solche im Buch enthalten sind, wie zum Beispiel „Wieviele Frauen, die jünger als 18 Jahre alt sind, arbeiteten in der Stadt Löwen im Textilsektor?". Komplexere Fragestellungen sind dagegen nur durch langwierige Suche und Nachrechnen per Hand zu beantworten, wie folgendes Beispiel verdeutlicht: „Was hatte größeren Einfluss auf den Arbeitslohn: Der entsprechende Industriesektor oder die Region?".

Heutzutage ermöglicht moderne Computertechnik den Forscherinnen und Forschern den Umgang mit großen Zahlenmengen. Dazu müssen gedruckte Texte bzw. Tabellen jedoch zunächst maschinenlesbar aufgearbeitet werden. Dies geschieht bei gedrucktem Material durch Digitalisierung und elektronische Texterkennung, sodass die Daten mithilfe von Computern berechnet bzw. bearbeitet werden können.

Elektronische Texterkennung für Fließtexte wird schon seit vielen Jahren betrieben. Die entsprechende Software ist mittlerweile sehr weit entwickelt. Die Umwandlung von Tabellen in maschinenlesbare Formate hat dagegen jedoch noch erhebliches Entwicklungspotenzial. Sie ist sinnvoll und notwendig, da es neben Volkszählungen noch unzählige andere Tabellen gibt, deren Digitalisierung erkenntnisreiche Ergebnisse hervorbringen kann (Wetterdaten, Börsendaten, Handel, Krankheitshäufigkeiten, ...).

Christian Clausner kommentiert die technischen Problemstellungen auf diesem Gebiet wie folgt:

8 Ministère de l'industrie et du travail. Office du travail - section de la statistique: Recensement general des industries et des metiers (31 octobre 1896). Bruxelles 1900-1902.

„*However, unlike narrative textual content, where digitisation is progressing well and in large scale, tabular numerical content is mostly untouched. The likely reason are the special challenges the processing of this kind of material poses: large quantity and complexity, low print / scan quality, variability of table layouts, changing content over time (for long time series), and requirements for very high accuracy.*" (Im Gegensatz zu narrativen Textinhalten, bei denen die Digitalisierung gut und in großem Maßstab voranschreitet, bleibt der tabellarische numerische Inhalt jedoch größtenteils unberührt. Der wahrscheinliche Grund sind die besonderen Herausforderungen, die die Verarbeitung dieser Art von Material mit sich bringt: große Menge und Komplexität, geringe Druck- und Scanqualität, wechselnde Tabellenlayouts, Änderung des Inhalts im Laufe der Zeit (für lange Zeitreihen) und Anforderungen an eine sehr hohe Genauigkeit.)[9]

Elektronische Texterkennung von Tabellen bringt demnach Herausforderungen bezüglich des Layouts und den hohen Anforderungen an verlässlich korrekte Erkennung mit sich. Dem Layout kommt hierbei eine große Bedeutung zu, da jede einzelne Ziffer ausschließlich in ihrem jeweiligen Kontext interpretierbar ist. Außerdem werden Fehler bei der Erkennung der Zahlen von den Nutzerinnen und Nutzern nicht in gleicher Weise wie gelegentliche Rechtschreibfehler in Fließtexten toleriert. Eine auf Wörterlisten basierende Fehlererkennung wie bei Fließtexten ist hier nicht brauchbar.

Für die digitale Aufbereitung von Tabellen unterscheidet Clausner zwischen vier Bearbeitungsniveaus:[10]

1. Die unterste Stufe bildet das Einscannen bzw. Abfotografieren der gedruckten Ausgabe. Hierdurch können Forscherinnen und Forscher jederzeit die Tabellen an ihrem PC anschauen und müssen nicht auf die wenigen noch verfügbaren Originalausgaben in Bibliotheken zurückgreifen.
2. Durch elektronische Texterkennung wird es Forscherinnen und Forschern ermöglicht, diese Dokumente maschinell zu durchsuchen.
3. Eine spezialisiertere Art der Texterkennung ermöglicht es, Tabellen in bearbeitbarer Form (z.B. Excel-Kalkulationstabellen) bereitzustellen. Hierdurch haben Forscherinnen und Forscher zum ersten Mal die Möglichkeit, bei der Analyse der Daten unmittelbar auf Software (z.B. Excel-Funktionalitäten) zurückgreifen zu können.
4. Die höchste Stufe besteht im Angebot der Dateien in einer Datenbank, sodass die Recherche übersichtlich ist und Berechnungen einfach und schnell durchgeführt werden können.

Beim Aufbau einer Datenbank in der mehrere verschiedene Zählungen kombiniert abgefragt werden können, gibt es weitere Herausforderungen zu bewältigen. Es muss beispielsweise darauf geachtet werden, dass die angewendeten Zählweisen identisch waren und die benutzten Kategorien dieselbe Bedeutung haben. So gilt es unter anderem zu klären, ob mit der Berufsbezeichnung Maler in jeder Tabelle ein Anstreicher oder in manchen Fällen nicht auch ein Künstler gemeint ist.

9 Clausner, Christian; Antonacopoulos, Apostolos; Henshaw, Christy u.a.: Towards the Extraction of Statistical Information from Digitised Numerical Tables: The Medical Officer of Health Reports Scoping Study, in: DATeCH2019: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, Brüssel, 2019, S. 65. Online: <https://doi.org/10.1145/3322905.3322932>.

10 Ebd., S. 66.

4. Digitalisierung von Tabellen an der KU Leuven

Im Zuge des Projekts der KU Leuven werden die Originalausgaben der Zählungen digitalisiert und die Texterkennung mittels OCR-Software durchgeführt. Das Endprodukt sind durchsuchbare Bilder und Excel-Kalkulationstabellen der gesamten Buchinhalte. Alle digitalisierten Zählungen sind im Bibliothekskatalog auffindbar oder auf der Website des Projektes https://bib.kuleuven.be/ebib/project-belgische-historische-tellingen/project_bht abrufbar. In den folgenden Absätzen wird unsere Arbeitsweise beschrieben.

4.1. Scannen / Fotografieren

Die Grundlage einer guten Texterkennung sind hochqualitative Bilder. Im Rahmen des Projekts der KU Leuven wird zwischen zwei Qualitätsstufen unterschieden, wobei erstere auf modernes Druckmaterial und letztere bei älteren Drucken angewendet wird.

Moderne Drucke, die nach 1900 entstanden sind, werden mit einer Auflösung von 400 dpi auf einem Canon imageFormula DR-G1100 Durchlaufscanner in der Fakultätsbibliothek für Wirtschaftswissenschaften gescannt. Dies ist jedoch nur dann durchführbar, wenn die gedruckten Originalausgaben der Zählungen zur Entnahme einzelner Seiten entbunden werden dürfen. Durch diese Methode kann eine mögliche Seitenwölbung verhindert werden, die die Texterkennung von Tabellen erheblich erschweren würde. Da die Zählungen fast ausschließlich aus Texten und Tabellen bestehen, wird bitonal, das heißt schwarz/weiß gescannt. Sollten in Ausnahmefällen farbige Karten enthalten sein, so werden diese gesondert gescannt und anschließend hinzugefügt.

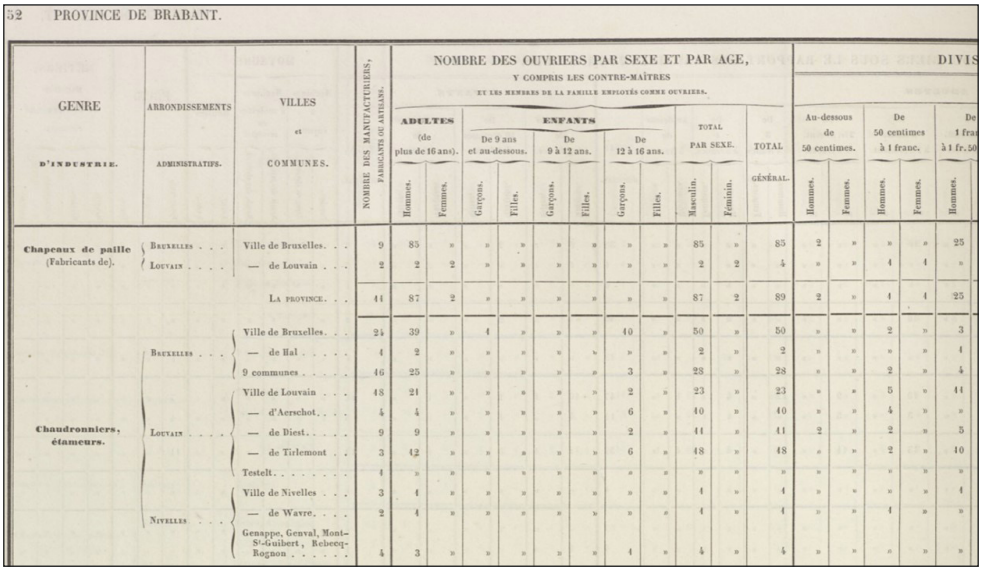
Die entbundenen Originalausgaben werden anschließend im Archiv der Fakultätsbibliothek aufbewahrt. Da sie dort nur auf Anfrage zugänglich sind, werden sie nicht wieder eingebunden. Die Benutzung dieser Bestände sollte vorzugsweise digital erfolgen. In vielen Fällen besitzt die KU Leuven jedoch mehrere Exemplare der Originalausgaben, sodass der Zugang auch zur gedruckten Form jederzeit gewährleistet ist.

Ältere Drucke oder Bücher, die nicht entbunden werden dürfen, werden durch den zentralen Digitalisierungsdienst der Universitätsbibliothek abfotografiert. Dieser Dienst verfügt über mehrere moderne technische Geräte zur Digitalisierung und benutzt je nach Zustand und Art der Buchbindung des Originals den bestmöglichen Scanner. In unserem Fall sind das entweder ein Quidenus Buchscanner oder Fotografie mithilfe eines Reprostandes. Beim Quidenus Buchscanner liegen die Bücher auf einer Buchwippe mit einem optimal eingestellten Öffnungswinkel, sodass die Buchbindung geschont wird. Eine Glasplatte, die auf die zu fotografierenden Seiten gelegt wird, minimiert die störende Wölbung. Beide Buchseiten werden anschließend durch zwei Nikon D850 Kameras fotografiert. Diese Arbeitsweise ist jedoch nur möglich, wenn der Bundsteg des Originaldokumentes breit genug ist, sodass der Rand der Glasplatte keine bedruckten Stellen verdecken und somit unlesbar machen würde.

Falls der Bundsteg zu schmal ist, werden die Bücher mithilfe eines Reprostandes fotografiert. Hierbei liegen die Bücher mit einer 180° Öffnung auf einer Fläche. Mit einer auf dem Reprostand befestigten Nikon D850 Kamera wird eine Aufnahme gemacht, die beide Buchseiten umfasst. Anschließend

müssen die Fotografien der Doppelseiten in ABBYY FineReader geteilt werden, sodass ein Bild einer Buchseite entspricht.

Die älteren Bücher werden in Farbe fotografiert. Diese Bilder werden den Nutzerinnen und Nutzern später zusätzlich zur Version mit Texterkennung und der Exceltabelle auch in dieser Farbaufnahme zur Verfügung gestellt.



52 PROVINCE DE BRABANT.

GENRE D'INDUSTRIE.	ARRONDISSEMENTS ADMINISTRATIFS.	VILLES et COMMUNES.	NOMBRE DES MANUFACTURIERS, FABRICANTS OU ARTISANS.	NOMBRE DES OUVRIERS PAR SEXE ET PAR AGE, Y COMPRIS LES CONTRE-MAÎTRES ET LES MEMBRES DE LA FAMILLE EMPLOYÉS COMME OUVRIERS.										DIVIS
				ADULTES (de plus de 16 ans).		ENFANTS De 9 ans et au-dessous.		De 9 à 12 ans.		De 12 à 16 ans.		TOTAL PAR SEXE.	TOTAL GÉNÉRAL.	Au-dessous de 50 centimes.
				Hommes.	Femmes.	Garçons.	Filles.	Garçons.	Filles.	Garçons.	Filles.	Masculin.	Féminin.		Hommes.	Femmes.
Chapeaux de paille (Fabricants de).	Bruxelles	Ville de Bruxelles.	9	85	»	»	»	»	»	»	»	85	»	85	2	»
	Louvain	— de Louvain	2	2	2	»	»	»	»	»	»	2	2	4	»	»
		La province.	11	87	2	»	»	»	»	»	»	87	2	89	2	»
Chaudronniers, étameurs.	Bruxelles	Ville de Bruxelles.	24	39	»	1	»	»	»	10	»	50	»	50	»	»
		— de Hal	1	2	»	»	»	»	»	»	»	2	»	2	»	»
		9 communes	16	25	»	»	»	»	»	3	»	28	»	28	»	»
	Louvain	Ville de Louvain	18	21	»	»	»	»	»	2	»	23	»	23	»	»
		— d'Aerschot.	4	4	»	»	»	»	»	6	»	10	»	10	»	»
		— de Diest.	9	9	»	»	»	»	»	2	»	11	»	11	2	»
		— de Tirlemont	3	12	»	»	»	»	»	6	»	18	»	18	»	»
		Testelt.	1	»	»	»	»	»	»	»	»	»	»	»	»	»
	Nivelles	Ville de Nivelles	3	1	»	»	»	»	»	»	»	1	»	1	»	»
		— de Wavre.	2	1	»	»	»	»	»	»	»	1	»	1	»	»
		Genappe, Genval, Mont-St-Guibert, Rebecq-Rognon	4	3	»	»	»	»	»	1	»	4	»	4	»	»

Abb. 1: Ausschnitt einer fotografierten Buchseite der Industriezählung von 1846 (eigene Bildschirmaufnahme)

4.2. Elektronische Text- und Zeichenerkennung

Die gescannten Seiten werden anschließend mit dem OCR-Programm ABBYY FineReader[11] weiter bearbeitet, wobei im ersten Schritt formale Verbesserungen der hochgeladenen Aufnahmen durchgeführt werden. Zunächst werden Schmutzflecken und der über die Buchseite hinausgehende Bereich („Schwarze Ränder" …) entfernt. Anschließend werden diese Aufnahmen so bearbeitet, dass sie für die Software lesbarer sind (Die Aufnahmen der eingescannten Buchseiten begradigen etc.). Daraufhin werden die drei Arbeitsschritte des Programmes durchlaufen: Analysieren, Texterkennen, Überprüfen.

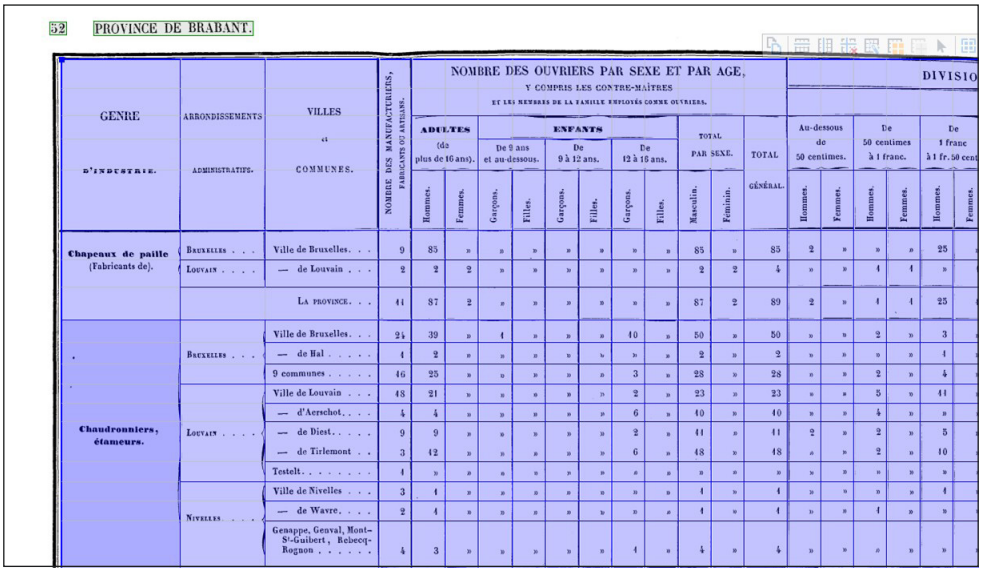
Beim Analysieren unterscheidet ABBYY FineReader zwischen Textbereichen, Tabellen und Bildern. Dies dient als Vorbereitung der späteren Erkennung. Bilder – in unserem Fall sind das Abbildungen des Logos des belgischen Staates oder Landkarten – werden nicht mehr weiter analysiert. Bei Texten wird die Textrichtung im Original definiert. Vor allem bei Tabellen kommt es vor, dass die Leserichtung von Textstellen variiert. In manchen Feldern verläuft sie von links nach rechts, in anderen Feldern

11 <https://www.abbyy.com/de-de/finereader>

von unten nach oben. Bei Tabellen werden die Positionen der jeweiligen Zahlen für eine später auf Grundlage des Images erstellte Excel-Tabelle durch vertikale und horizontale Linien definiert.

In der Praxis hat sich jedoch gezeigt, dass die beiden letztgenannten Punkte bei komplexen Tabellen oder Seiten, die etwas schräg eingescannt wurden, noch nicht einwandfrei funktionieren. Die betroffenen Textstellen bzw. Tabellen werden in diesem Falle händisch nachbearbeitet. Diese händische Nachbearbeitung ist neben der Endredaktion in Excel die zeitaufwendigste Aufgabe des gesamten Arbeitsprozesses. ABBYY FineReader bietet jedoch genügend Werkzeuge an, um eine korrektes Tabellenlayout zu erstellen.

Einerseits kann das Optimalisieren des Bildes (Schräglagekorrektur, Korrektur von Trapezverzerrungen, …) die Strukturerkennung begünstigen, andererseits ermöglicht ABBYY FineReader die Bearbeitung der erkannten Tabellenstruktur durch hinzufügen, entfernen oder verschieben von Linien sowie durch das Verbinden und Teilen von Zellen. Außerdem muss auch darauf geachtet werden, dass die Leserichtung jeder Textstelle richtig erfasst wurde. Diese kann in ABBYY FineReader bei Bedarf für jeden Bereich manuell eingestellt werden.

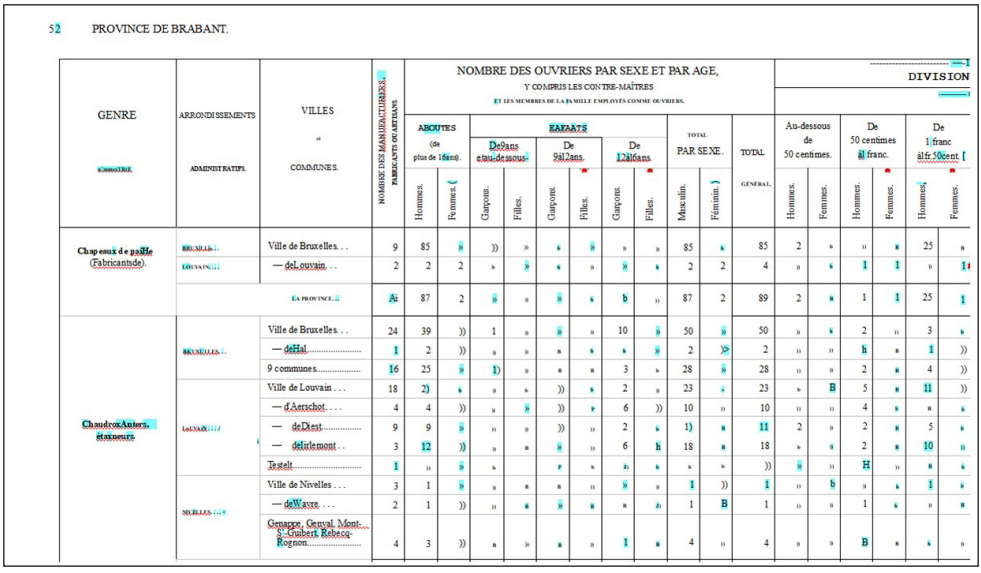


52 PROVINCE DE BRABANT.

GENRE D'INDUSTRIE.	ARRONDISSEMENTS ADMINISTRATIFS.	VILLES et COMMUNES.	NOMBRE DES MANUFACTURIERS, FABRICANTS OU ARTISANS.	NOMBRE DES OUVRIERS PAR SEXE ET PAR AGE, Y COMPRIS LES CONTRE-MAÎTRES ET LES MEMBRES DE LA FAMILLE EMPLOYÉS COMME OUVRIERS.										DIVISIO					
				ADULTES (de plus de 16 ans).		ENFANTS De 9 ans et au-dessous.		De 9 à 12 ans.		De 12 à 15 ans.		TOTAL PAR SEXE.		TOTAL GÉNÉRAL.	Au-dessous de 50 centimes.		De 50 centimes à 1 franc.		De 1 franc à 1 fr. 50 cent
				Hommes.	Femmes.	Garçons.	Filles.	Garçons.	Filles.	Garçons.	Filles.	Masculin.	Féminin.		Hommes.	Femmes.	Hommes.	Femmes.	Hommes.	Femmes.
Chapeaux de paille (Fabricants de).	Bruxelles	Ville de Bruxelles.	9	85	»	»	»	»	»	»	»	85	»	85	2	»	»	»	25	
	Louvain	— de Louvain	2	2	2	»	»	»	»	»	»	2	2	4	»	»	1	1	»	
		La province.	11	87	2	»	»	»	»	»	»	87	2	89	2	»	1	1	25	
Chaudronniers, étameurs.	Bruxelles	Ville de Bruxelles.	24	39	»	1	»	»	»	10	»	50	»	50	»	»	2	»	3	
		— de Hal	1	2	»	»	»	»	»	»	»	2	»	2	»	»	»	»	1	
		9 communes	16	25	»	»	»	»	»	3	»	28	»	28	»	»	2	»	4	
	Louvain	Ville de Louvain	18	21	»	»	»	»	»	2	»	23	»	23	»	»	5	»	11	
		— d'Aerschot.	4	4	»	»	»	»	»	6	»	10	»	10	»	»	4	»	»	
		— de Diest.	9	9	»	»	»	»	»	2	»	11	»	11	2	»	2	»	5	
		— de Tirlemont	3	12	»	»	»	»	»	6	»	18	»	18	»	»	2	»	10	
		Testelt.	1	»	»	»	»	»	»	»	»	»	»	»	»	»	»	»	»	
	Nivelles	Ville de Nivelles	3	1	»	»	»	»	»	»	»	1	»	1	»	»	»	»	1	
		— de Wavre.	2	1	»	»	»	»	»	»	»	1	»	1	»	»	1	»	»	
		Genappe, Genval, Mont-St-Guibert, Rebecq-Rognon	4	3	»	»	»	»	»	1	»	4	»	4	»	»	»	»	»	

Abb. 2: Ausschnitt der gleichen Seite (s. Abb. 1) nach der Analyse in ABBYY FineReader (eigene Bildschirmaufnahme der Anwendung ABBYY FineReader)

Im Anschluss an diesen vorbereitenden Arbeitsschritt führt die Software die eigentliche Texterkennung durch. Neben dem Fenster mit den Bildern erscheint ein weiteres Fenster mit dem OCR-Text.



52 PROVINCE DE BRABANT.

GENRE	ARRONDISSEMENTS ADMINISTRATIFS.	VILLES et COMMUNES.	NOMBRE DES MANUFACTURES, FABRICANTS OU ARTISANS.	NOMBRE DES OUVRIERS PAR SEXE ET PAR AGE, Y COMPRIS LES CONTRE-MAÎTRES ET LES MEMBRES DE LA FAMILLE EMPLOYÉS COMME OUVRIERS.										DIVISION	
				ADOULTES (de plus de 16ans).		ENFANTS				TOTAL PAR SEXE.		TOTAL GÉNÉRAL.	Au-dessous de 50 centimes.		De 50 centimes à1 franc.		De 1 franc à1fr.50cent.	
						De9ans a moins de9ans.		De 9à12ans.		De 12à16ans.								
				Hommes.	Femmes.	Garçons.	Filles.	Garçons.	Filles.	Garçons.	Filles.	Masculin.	Féminin.		Hommes.	Femmes.	Hommes.	Femmes.	Hommes.	Femmes.
Chapeaux de paille (Fabricants de).	BRUXELLES.	Ville de Bruxelles. . .	9	85	»	))	»	»	»	»	»	85	»	85	2	»	»	»	25	»
	LOUVAIN.	— de Louvain. . .	2	2	2	»	»	»	»	»	»	2	2	4	»	»	1	1	»	1
		EA PROVINCE.	Ai	87	2	»	»	»	»	b	»	87	2	89	2	»	1	1	25	1
Chaudronniers, étameurs.	BRUXELLES.	Ville de Bruxelles. . .	24	39	))	1	»	»	»	10	»	50	»	50	»	»	2	»	3	»
		— de Hal.	1	2	))	»	»	»	»	»	»	2	))	2	»	»	h	»	1	))
		9 communes.	16	25	»	1)	»	»	»	3	»	28	»	28	»	»	2	»	4	))
	LOUVAIN.	Ville de Louvain . . .	18	2)	»	»	»	))	»	2	»	23	»	23	»	B	5	»	11	))
		— d'Aerschot. . . .	4	4	))	»	»	))	»	6	))	10	»	10	»	»	4	»	»	»
		— de Diest.	9	9	»	»	»	))	»	2	»	1)	»	11	2	»	2	»	5	»
		— de Wavre.	3	12	))	»	»	»	»	6	h	18	»	18	»	»	2	»	10	»
		Testelt.	1	»	»	»	»	»	»	»	»	»	»	))	»	»	H	»	»	»
	NIVELLES.	Ville de Nivelles . . .	3	1	»	»	»	»	»	»	1	))	1	»	b	»	»	1	»
		— de Wavre. . . .	2	1	))	»	»	»	»	»	1	B	1	»	»	1	»	»	»
		Genappe, Genval, Mont-St-Guibert, Rebecq-Rognon.	4	3	))	»	»	»	»	1	»	4	»	4	»	»	B	»	»	»

Abb. 3: Ergebnis nach der Texterkennung desselben Ausschnittes (eigene Bildschirmaufnahme der Anwendung ABBYY FineReader)

Im letzten Arbeitschritt in ABBYY FineReader werden unzuverlässig erkannte Zeichen überprüft. Hierbei werden alle Zeichen in einem zweigeteilten Dialogfenster durchlaufen. Im oberen Teil wird das Bild der Zeichengruppe gezeigt, die das Programm möglicherweise nicht korrekt erkannt hat, und im unteren Teil der Vorschlag, den man gegebenenfalls verbessern kann.

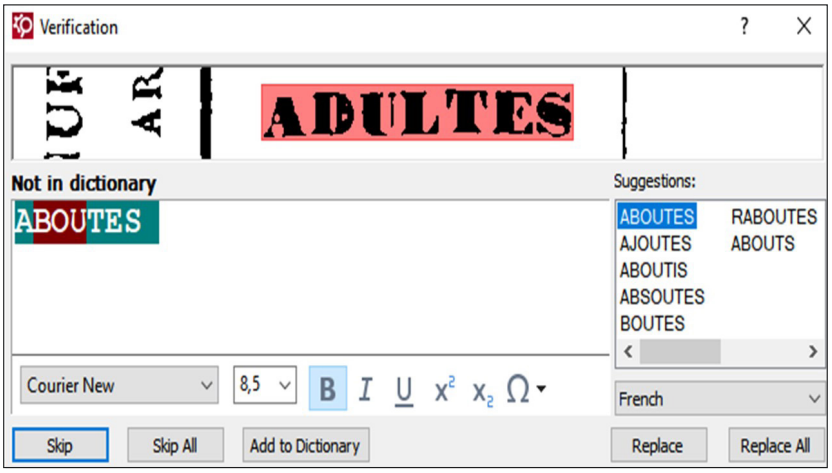


Abb. 4: Überprüfung (eigene Bildschirmaufnahme der Anwendung ABBYY FineReader)

Beim Überprüfen greift das Programm auf Wörterbücher der im Originaltext benutzten Sprache(n) zurück. Problematischer ist es bei Ziffern, da in diesem Fall eine wörterbuchgestützte Korrektur unmöglich ist. Ziffern kommen auch nicht immer in der Überprüfung vor, deshalb werden im späteren Verlauf noch ein weiteres Mal ausschließlich die Ziffern in den Tabellen überprüft.

Vorher wird das Dokument als durchsuchbares pdf und als Excel-Datei abgespeichert. Das pdf-Dokument wird nicht mehr weiter behandelt, da es bereits in diesem Stadium ausreichend durchsuchbar ist. Mit der Excel-Tabelle verhält es sich anders – diese muss noch weiter bearbeitet werden.

Die Zählungen bestehen meistens aus einigen wenigen Tabellen, die sich auf hunderte Seiten erstrecken können. Deshalb wird der Inhalt in der Excel Tabelle auf mehrere Arbeitsblätter verteilt, sodass jede umfangreiche Tabelle auf einem gesonderten Arbeitsblatt gespeichert wird.

Beim Übertragen der OCR-Daten in Excel erhält man als Resultat eine vertikale Aneinanderreihung aller Seiten der gedruckten Ausgabe. Diese Seiten müssen nun wieder logisch zusammengefügt werden, wobei die Tabellen nicht nur vertikal, sondern auch horizontal aneinandergefügt werden müssen. Im einfachsten Fall ist die Tabelle eine Buchseite breit. Dann werden die in Excel übertragenen Daten der Buchseite einfach untereinander zusammengefügt. Eine Tabelle kann jedoch auch eine Dopelseite breit sein, sodass Seite 2 rechts neben Seite 1, und die Seiten 3 und 4 unter den Seiten 1 und 2 zusammengefügt werden.

Meistens bestehen die Tabellen jedoch aus so vielen Spalten, dass vier Seiten horizontal aneinandergereit werden, die nächsten vier Seiten darunter usw.

In der folgenden Abbildung werden einige mögliche Buchseitenanordnungen dargestellt, wie sie in Excel übernommen werden sollten, damit die Tabellenstruktur bewahrt wird.

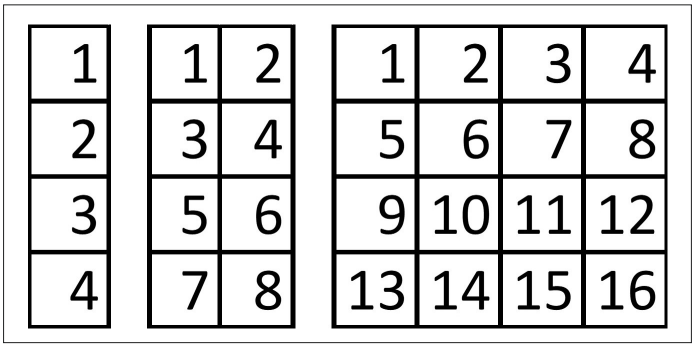


Abb. 5: Mögliche Seitenanordnungen

Zur besseren Orientierung wird in die Excel-Datei eine Spalte mit den Seitenzahlen des Originaldokumentes hinzugefügt. Dies vereinfacht zudem das parallele Arbeiten mit den verschiedenen Wiedergaben der Tabellen (Excel, pdf oder die gedruckte Originalausgabe).

Anschließend sind noch weitere Veränderungen des Layouts erforderlich, mit dem Ziel die Klarheit der Tabellenstruktur zu steigern. Diese Anpassungen sind vor allem ästhetischer Natur.

Im letzten Arbeitsschritt erfolgt eine weitere Kontrolle der Ziffern mithilfe von Excel-Funktionalitäten. Diese Kontrolle stützt sich auf inhaltliche Berechnungen mehrerer Zeilen und Spalten.

Zur Verdeutlichung sei hier folgendes Beispiel genannt: In Reihe 3 werden Brauereien gezählt. Spalte C gibt Brauereien mit 1-5 Mitarbeitern an, Spalte D mit 6-10 Mitarbeitern, Spalte E 11-20 Mitarbeiter, usw. In Spalte H steht letzendlich die Gesamtzahl aller Brauereien. Die Kontrolle wird ausgeführt, indem die berechnete Gesamtzahl der Spalten C3, D3, E3, F3 und G3 mit dem Wert der Spalte H3 verglichen wird. Dazu wird eine Spalte I hinzugefügt mit der Excel-Formel: =SUM(C3:G3)=H3 . Wenn die Berechnung stimmt, dann erscheint ein WAHR (TRUE), ansonsten ein FALSCH (FALSE). Diese Formel wird auf die ganze Spalte angewandt, sodass Reihen mit fehlerhaften Ziffern schnell ausfindig gemacht werden können.

Im Beispiel ergibt die Summe der Mitarbeiter jeder Altersklasse in Brauereien die Gesamtzahl 10. Dies entspricht auch der Gesamtsumme im Originaldokument. Bei den Mitarbeitern in Molkereien ergibt sich jedoch eine errechnete Summe von 15 Mitarbeitern, obwohl die Gesamtsumme 16 sein sollte. Anhand des Originaldokumentes wird die fehlerhaft übertragene Ziffer ausfindig gemacht und anschließend in der Excel-Tabelle korrigiert.

	Anzahl Mitarbeiter						
	1 bis 5	6 bis 10	11 bis 20	21 bis 50	50 und mehr	Gesamt	
Brauereien	3	1	2	3	1	10	TRUE
Molkereien	0	7	3	3	2	16	FALSE

Abb. 6: Beispiel einer Berechnung

4.3. Auffindbarkeit und Präsentation der Daten

Nachdem alle Arbeitsschritte zur Korrektur der Texterkennung erledigt sind, werden die hochauflösenden Fotos, das durchsuchbare PDF-Dokument und die Excel-Kalkulationstabelle in Teneo[12] abgelegt. Die hochauflösenden Fotos werden in den drei Formaten TIFF, JPEG und JPEG2000 gespeichert, wobei die TIFF-Dateien ausschließlich der Langzeitarchivierung dienen und nicht für die Öffentlichkeit verfügbar sind. Die JPEG, JPEG2000, PDF und Excel-Dateien sind über den Bibliothekskatalog Limo und via Google auffindbar und werden mit dem in Rosetta integriertem General IE viewer als Open Data[13] angeboten. Sie stehen Interessierten somit zur freien Nutzung zur Verfügung.

12 Teneo <http://www.libis.be/teneo> ist eine Erweiterung von Rosetta (Ex Libris) zur Langzeitbewahrung und wird an der KU Leuven als Produktname hierfür gebraucht.

13 KU Leuven Libraries: Images as open data, 13.03.2020, <https://bib.kuleuven.be/english/BD/digit/digitisation/images-as-open-data>, Stand: 13.03.2020.

Die Industriezählung von 1846 eignet sich sehr gut als Beispiel, um diese Vorgehensweise zu veranschaulichen.[14]

5. Herausforderungen und Ausblick

Die Digitalisierung von Tabellen ist heutzutage noch sehr zeitaufwendig, doch durch technologische Weiterentwicklung könnte es in einigen Jahren möglich sein, mit weniger Aufwand befriedigende Resultate zu erzielen.

Ist die Qualitätskontrolle bei der Erfassung textueller Dokumente noch verhältnismäßig einfach, da sich das OCR-Programm auf den Vergleich mit Wörterlisten stützen kann, so sind einzelne Ziffern schwieriger auf ihre Richtigkeit zu überprüfen. Bei älteren Dokumenten eignet sich außerdem der Schrifttyp nicht immer für ein korrektes Lesen. Häufige Fehler sind unter anderem das Verwechseln der Ziffern 1 und 4 sowie 5 und S. Die Lösung die an der KU Leuven angewandt wird, umfasst eine Endkontrolle durch Berechnungen der Werte in Excel. Idealerweise erfolgt dies mit vertikalen und horizontalen Berechnungen. Nur so können nahezu alle fehlerhaften Übertragungen ausgeschlossen werden.

Mit dieser Methode können übrigens auch fehlerhafte Angaben im ursprünglichen Dokument erfasst werden. Wird die fehlerhafte Angabe eindeutig identifiziert und die Richtigstellung kann durch eine vertikale und horizontale Berechnung doppelt bestätigt werden, dann wird die fehlerhafte Zahl durch die wahrscheinlichere ersetzt und diese Zelle wird mit blauer Farbe markiert. Falls fehlerhafte Angaben zwar identifiziert werden können, die Richtigstellung jedoch nicht eindeutig ist, dann werden diese Zellen mit roter Farbe markiert.

Der Aufbau von Datenbanken zum Konsultieren von Tabellen verschiedener Jahrgänge bringt gleich mehrere inhaltliche Probleme mit sich. So wäre zum einen das Problem, dass nicht alle Kategorien in jeder Zählung die gleiche Bedeutung haben. Zum anderen muss ebenfalls darauf geachtet werden, dass auch bei gleicher Bedeutung die Zählweise identisch ist. Da innerhalb des heutigen Projektes eine Datenbank nicht vorgesehen ist, wird sich damit nicht weiter befasst. Diese Studien gehören eher zum Aufgabenbereich der interessierten Forscherinnen und Forscher.

Die Bibliothek für Wirtschaftswissenschaften der KU Leuven beschränkt sich vorerst auf die Übertragung aller Industriezählungen in durchsuchbare pdf-Dokumente und bearbeitbare Excel-Tabellen. Diese werden für jeden Interessierten frei zur Verfügung gestellt. Bei der Texterkennung werden Prozesse ständig weiterentwickelt. Ziel ist es, nach der Digitalisierung der Industriezählungen auch die belgischen Landwirtschafts- und Volkszählungen auf diese Weise aufzubereiten.

14 Statistique de la Belgique: Industrie, recensement général (15 octobre 1846). Online: <http://resolver.libis.be/IE11452503/representation> (Bild) und <http://resolver.libis.be/IE13011283/representation> (Durchsuchbarer Text + Excel-Kalkulationstabelle).

Literaturverzeichnis

- Bracke, Nele: Een monument voor het land. Overheidsstatistiek in België 1795-1870, Gent 2008. Online: <https://www.oapen.org/search?identifier=366390> (Stand: 29.04.2020).

- Černý, Jaroslav: Consanguineous Marriages in Pharaonic Egypt, in: Journal of Egyptian Archeology 40, 1954, S. 28–29. Online: <https://www.jstor.org/stable/3855544> (Stand: 29.04.2020).

- Clausner, Christian; Antonacopoulos, Apostolos; Henshaw, Christy u.a.: Towards the Extraction of Statistical Information from Digitised Numerical Tables: The Medical Officer of Health Reports Scoping Study, in: DATeCH2019: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, Brüssel, 2019, S. 65-71. Online: <https://doi.org/10.1145/3322905.3322932> (Stand: 29.04.2020).

- Klep, Paul: Politieke strubbelingen rond de volkstelling 1859, Voorburg, 2007. Online: <http://www.volkstelling.nl/nl/documentatie/1859/rede_pklep/index.html> (Stand: 29.04.2020).

- March, Lucien: Le recensement des industries en Belgique en 1896, in: Journal de la société statistique de Paris 43, 1902, S. 257-267.

- Project Belgische historische tellingen. Online: <https://bib.kuleuven.be/ebib/project-belgische-historische-tellingen/project_bht> (Stand: 29.04.2020).

- Statistique de la Belgique: Industrie, recensement général (15 octobre 1846). Online: <http://resolver.libis.be/IE11452503/representation> (Bild) und <http://resolver.libis.be/IE13011283/representation> (Durchsuchbarer Text + Excel-Kalkulationstabelle)

- Tenney, Frank: Roman Census Statistics from 508 to 225 B.C., in: The American Journal of Philology 51 (4), 1930, S. 313–324. Online: <https://www.jstor.org/stable/289892> (Stand: 29.04.2020).