University of Salford MANCHESTER
1967-2017 50 YEARS

Census 1961

Digitalisierung von historischen Zensus-Tabellen

Christian Clausner

28.09.2017

Projekt

- Census 1961 for England and Wales
- Pilotprojekt + zwei Verlängerungen (2016/2017)
- In Zusammenarbeit mit UK Data Service
- Teilweise finanziert von ONS (Office for National Statistics)
- Größeres Projekt geplant Ende 2017 (ein Jahr)
- Ähnliche Projekte evtl. später (Nordirland, Schottland…)

Dokumente

HAMMERSMITH MET.B. C.D.229/A E.D.001 NO. 1 COLLEGE PARK AND LATIMER 2

DISTRIBUTION BY TENURE

	HSDS	PSNS	ROOMS
OWNER-OCCUPIERS	58	188	254
RENTING W. BUSINESS	2	8	10
HOLDING BY EMPLMENT	4	12	16
RENTING FRM COUNCIL	3	15	7
RENTING FURNISHED	13	37	40
RENTING UNFURNISHED	151	408	494

	DWELLS	HSDS	PSNS	ROOMS
BDG TYPE I	58	72	238	332
BDG TYPE II	16	16	44	79
BDG TYPE III	146	147	386	438

RESIDENTS BORN OUTSIDE U.K.

BIRTHPLACES	MALE	FMLE
IRELAND	30	23
INDIA, PAKISTAN, CEYLON	-	-
BRIT.W.AFRICA	-	-
BR.E.&C.AFRICA	-	-
BR.CARIB.TERRS.	7	7
MALTA	-	-
CYPRUS	3	4
OTHER COMM.AREAS	2	3
FOREIGN AREAS	5	6
NATIONALITIES		
U.K.CITIZENS	14	14
OTHER BRITISH	-	1
EUROPEAN NATS.	3	4
OTHER ALIENS	-	1

NON-PRIVATE POPULATION - NOT HOTELS

	ALL PERSONS MALE	ALL PERSONS FMLE	INMATES MALE	INMATES FMLE	NO.OF ESTS
NHS ACUTE HOSPLS	-	-	-	-	-
NHS CHRONIC HOSP	-	-	-	-	-
NHS PSYCHIATRIC	-	-	-	-	-
NHS ISOLATION HOS	-	-	-	-	-
NHS OTHER HOSPS	-	-	-	-	-
OTHER MATERNITY	-	-	-	-	-
OTHER PSYCHIATRIC	-	-	-	-	-
OTHER CONVALESCEN	-	-	-	-	-
OTHER HOSPITALS	-	-	-	-	-
HOMES FOR AGED	-	-	-	-	-
HOMES FOR DISABLD	-	-	-	-	-
AGED AND DISABLED	-	-	-	-	-
CHILDRENS HOMES	-	-	-	-	-
EDUCATIONAL ESTBS	-	-			-
PCES OF DETENTION	-	-	-	-	-
DEFENCE ESTABS	-	-			-
CIVILIAN VESSELS	-	-			-
MISCEL. COMMUNAL	-	-			-
MISCELLANEOUS	-	-			-

HOUSEHOLD ARRANGEMENTS

		ALL HSDS	SHG HSDS	SHG KTCH
COLD WATR	SHRD	8	6	4
	NONE	-	-	-
HOT WATER	SHRD	4	1	1
	NONE	155	27	6
FIXD BATH	SHRD	16	5	1
	NONE	173	26	6
WATR CLST	SHRD	126	28	7
	NONE	4	1	-
ALL EXCLUSIVE		30	1	-

OLD PERSONS ALONE 3 M. 17 F.

OLD PERSONS IN HSDS OF 2PERSONS-

	ONE OLD	TWO OLD
MARRIED COUPLE	7	9
OTHERS- MALE OLDER	2	1
FEMALE OLDER	1	-
BOTH MALE	-	-
BOTH FEMALE	1	1

HOTELS OF UNDER 10 ROOMS

NO.OF HOTELS	1
TOTAL ROOMS	6
PERSONS ENUM.	12
HOTELS OF 10 OR MORE RMS	
10-14 RMS	-
15-24 RMS	-
25-49 RMS	-
50-99 RMS	-
100-199	-
200 OR MORE	-
TOTAL RMS	-
PERSONS ENUMERATED	
MANAGER AND STAFF	-
RELATIVES OF STFF	-
RESIDENT GUESTS	-
VISITOR GUESTS	-

ALL NON-PRIVATE

	MALE	FMLE
0-	1	-
5-	-	-
10-	-	-
15-	2	1
20-	-	3
25-	3	-
30-	-	-
35-	-	-
40-	-	-
45-	-	-
50-	-	-
55-	-	-
60-	-	2
65-	-	-

DENSITIES - PERSONS PER ROOM

	OVR1.5	1-1.5	1	0.75-	.5-	UND.5
ALL HSDS	21	29	47	12	89	33
WTH KTCH	2	3	3	4	11	2
SHG KTCH	2	2	3	-	-	-
PERSONS	108	119	149	47	202	43

PERSONS RESIDENT OUTSIDE L.A.AREA

	MALE	FMLE
REST OF ENG/WAL	2	-
OUTSIDE ENG/WALES	-	-

AGE AND MARITAL CONDITION BY FIVE YEAR AGE GROUPS

	PERSONS	MALES TOTAL	MALES SINGLE	MALES MARRID	MALES WIDWD	MALES DVD	FEMALES TOTAL	FEMALES SINGLE	FEMALES MARRID	FEMALES WIDWD	FEMALES DVD
TOTAL	680	320	132	179	8	1	360	136	176	46	2
0- 4	63	36	36				27	27			
5- 9	47	24	24				23	23			
10-14	36	13	13				23	23			
15-19	50	22	20	2	-	-	28	24	4	-	-
20-24	50	23	14	9	-	-	27	9	18	-	-
25-29	47	26	7	19	-	-	21	4	17	-	-
30-34	40	21	2	19	-	-	19	3	16	-	-
35-39	37	18	3	15	-	-	19	2	16	-	1
40-44	46	19	3	15	-	1	27	5	21	-	1
45-49	68	34	2	32	-	-	34	3	28	3	-
50-54	61	30	4	26	-	-	31	5	22	4	-
55-59	37	18	2	16	-	-	19	3	12	4	-
60-64	37	18	2	13	3	-	19	1	11	7	-
65-69	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
70-74	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
75-79	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
80-84	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
85-89	1	-	-	-	-	-	1	-	-	1	-
90-94	3	-	-	-	-	-	3	-	-	3	-
95-	-	-	-	-	-	-	-	-	-	-	-

SINGLE YEARS UNDER 21

AGES	PERSONS	MALES	FMLES
0-20	202	97	105
0	19	13	6
1	7	4	3
2	16	4	12
3	11	8	3
4	10	7	3
5	13	6	7
6	10	5	5
7	7	3	4
8	7	5	2
9	10	5	5
10	3	3	-
11	4	1	3
12	7	1	6
13	[illegible]	[illegible]	[illegible]
14	[illegible]	[illegible]	[illegible]
15	[illegible]	[illegible]	[illegible]
16	[illegible]	[illegible]	[illegible]
17	6	3	3
18	10	5	5
19	9	2	7
20	6	2	4

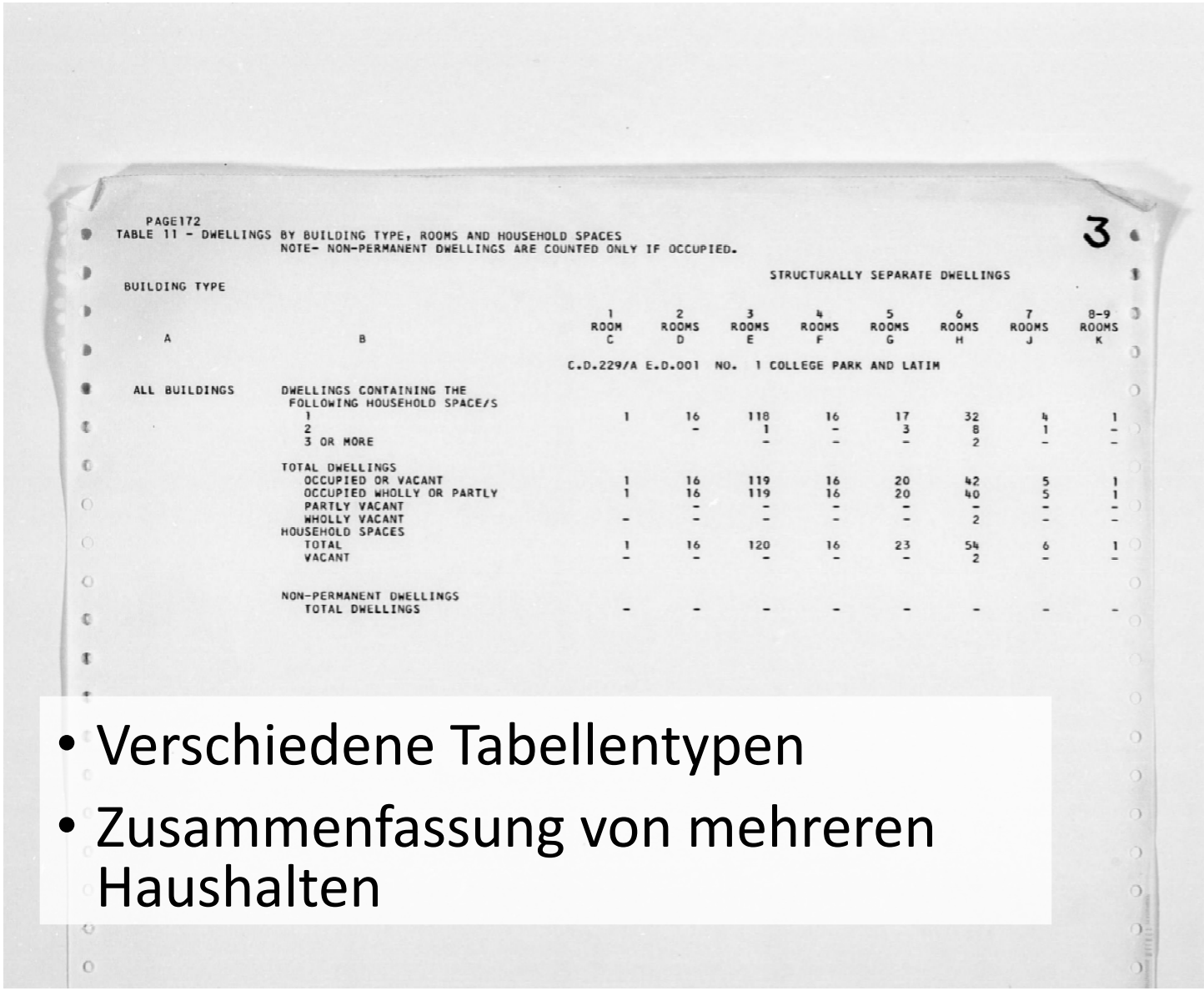
SHARING HOUSEHOLDS BY NO. OF PERSONS

1	9
2	10
3	8
4	2
5	2
6	-
7-	1

TENURE	HSDS
	58
ESS	2
ENT	4
CIL	3
D	13
HED	151

- Großteil „Small Area Statistics“ - SAS

Dokumente



PAGE172
TABLE 11 - DWELLINGS BY BUILDING TYPE, ROOMS AND HOUSEHOLD SPACES
NOTE- NON-PERMANENT DWELLINGS ARE COUNTED ONLY IF OCCUPIED.

3

BUILDING TYPE		STRUCTURALLY SEPARATE DWELLINGS							
A	B	1 ROOM C	2 ROOMS D	3 ROOMS E	4 ROOMS F	5 ROOMS G	6 ROOMS H	7 ROOMS J	8-9 ROOMS K
		C.D.229/A E.D.001 NO. 1 COLLEGE PARK AND LATIM							
ALL BUILDINGS	DWELLINGS CONTAINING THE FOLLOWING HOUSEHOLD SPACE/S								
	1	1	16	118	16	17	32	4	1
	2		-	1	-	3	8	1	-
	3 OR MORE			-	-	-	2	-	-
	TOTAL DWELLINGS								
	OCCUPIED OR VACANT	1	16	119	16	20	42	5	1
	OCCUPIED WHOLLY OR PARTLY	1	16	119	16	20	40	5	1
	PARTLY VACANT		-	-	-	-	-	-	-
	WHOLLY VACANT	-	-	-	-	-	2	-	-
	HOUSEHOLD SPACES								
	TOTAL	1	16	120	16	23	54	6	1
	VACANT	-	-	-	-	-	2	-	-
	NON-PERMANENT DWELLINGS								
	TOTAL DWELLINGS	-	-	-	-	-	-	-	-

- Verschiedene Tabellentypen
- Zusammenfassung von mehreren Haushalten

Dokumente

PAGE 172

10 OR MORE ROOMS L	TOTAL M	PARTLY VACANT N	VACANT O	NUMBER OF ROOMS TOTAL P	OCCUPIED Q	VACANT R
-	205		2	764	752	12
-	13	-	-	73	73	-
-	2	-	-	12	12	-
-	220	-	2	849	837	12
-	218	-		837	837	-
-	-	-		-	-	-
-	2		2	12		12
-	237	-				
-	2					
-	-	-		-	-	-

- Originale sind Computerausdrucke
- Ca. 100.000 Scans von Mikrofilm

Dokumente

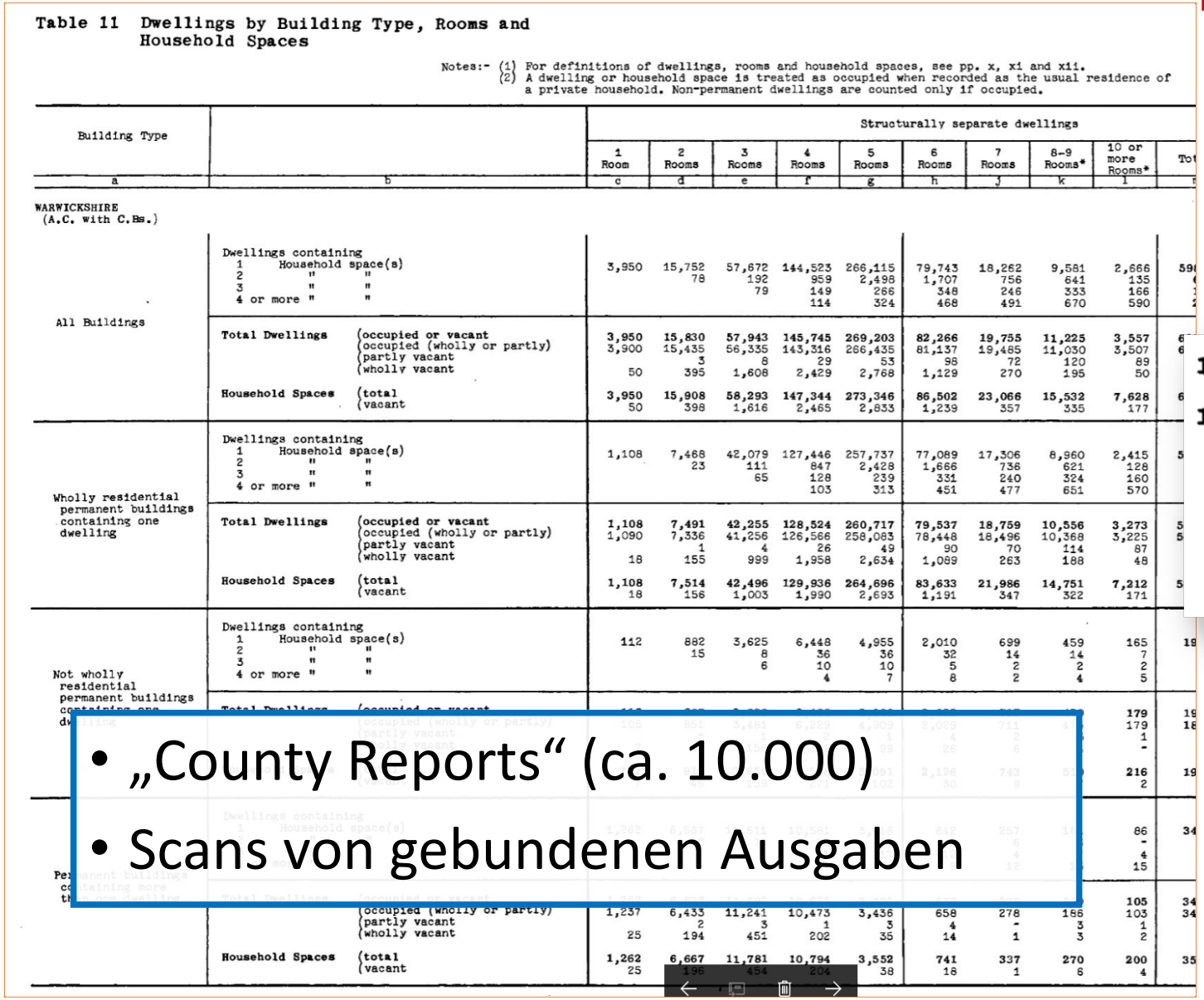


Table 11 Dwellings by Building Type, Rooms and Household Spaces

Notes:- (1) For definitions of dwellings, rooms and household spaces, see pp. x, xi and xii.
(2) A dwelling or household space is treated as occupied when recorded as the usual residence of a private household. Non-permanent dwellings are counted only if occupied.

Building Type		Structurally separate dwellings: 1 Room	2 Rooms	3 Rooms	4 Rooms	5 Rooms	6 Rooms	7 Rooms	8-9 Rooms*	10 or more Rooms*	Tot
a	b	c	d	e	f	g	h	j	k	l	
WARWICKSHIRE (A.C. with C.Bs.)											
All Buildings	Dwellings containing 1 Household space(s)	3,950	15,752	57,672	144,523	266,115	79,743	18,262	9,581	2,666	59
	Dwellings containing 2 " "		78	192	959	2,498	1,707	756	641	135	
	Dwellings containing 3 " "			79	149	266	348	246	333	166	
	Dwellings containing 4 or more " "				114	324	468	491	670	590	
	Total Dwellings (occupied or vacant	3,950	15,830	57,943	145,745	269,203	82,266	19,755	11,225	3,557	
	Total Dwellings (occupied (wholly or partly)	3,900	15,435	56,335	143,316	266,435	81,137	19,485	11,030	3,507	
	Total Dwellings (partly vacant		3	8	29	53	98	72	120	89	
	Total Dwellings (wholly vacant	50	395	1,608	2,429	2,768	1,129	270	195	50	
	Household Spaces (total	3,950	15,908	58,293	147,344	273,346	86,502	23,066	15,532	7,628	
	Household Spaces (vacant	50	398	1,616	2,465	2,833	1,239	357	335	177	
Wholly residential permanent buildings containing one dwelling	Dwellings containing 1 Household space(s)	1,108	7,468	42,079	127,446	257,737	77,089	17,306	8,960	2,415	
	Dwellings containing 2 " "		23	111	847	2,428	1,666	736	621	128	
	Dwellings containing 3 " "			65	128	239	331	240	324	160	
	Dwellings containing 4 or more " "				103	313	451	477	651	570	
	Total Dwellings (occupied or vacant	1,108	7,491	42,255	128,524	260,717	79,537	18,759	10,556	3,273	
	Total Dwellings (occupied (wholly or partly)	1,090	7,336	41,256	126,566	258,083	78,448	18,496	10,368	3,225	
	Total Dwellings (partly vacant		1	4	26	49	90	70	114	87	
	Total Dwellings (wholly vacant	18	155	999	1,958	2,634	1,089	263	188	48	
	Household Spaces (total	1,108	7,514	42,496	129,936	264,696	83,633	21,986	14,751	7,212	
	Household Spaces (vacant	18	156	1,003	1,990	2,693	1,191	347	322	171	
Not wholly residential permanent buildings containing one dwelling	Dwellings containing 1 Household space(s)	112	882	3,625	6,448	4,955	2,010	699	459	165	
	Dwellings containing 2 " "		15	8	36	36	32	14	14	7	
	Dwellings containing 3 " "			6	10	10	5	2	2	2	
	Dwellings containing 4 or more " "				4	7	8	2	4	5	
	Total Dwellings (occupied or vacant									179	
	Total Dwellings (occupied (wholly or partly)									179	
	Total Dwellings (partly vacant									1	
	Total Dwellings (wholly vacant									-	
	Household Spaces (total									216	
	Household Spaces (vacant									2	
Permanent buildings containing more than one dwelling	Dwellings containing 1 Household space(s)									86	
	Dwellings containing 2 " "									-	
	Dwellings containing 3 " "									4	
	Dwellings containing 4 or more " "									15	
	Total Dwellings (occupied or vacant									105	
	Total Dwellings (occupied (wholly or partly)	1,237	6,433	11,241	10,473	3,436	658	278	166	103	
	Total Dwellings (partly vacant		2	3	1	3	4	-	3	1	
	Total Dwellings (wholly vacant	25	194	451	202	35	14	1	3	2	
	Household Spaces (total	1,262	6,667	11,781	10,794	3,552	741	337	270	200	
	Household Spaces (vacant	25	196	454	204	38	18	1	6	4	

190,288	1·3
129,293	8·4
60,995	0·5
1,955	2·0
32,245	12·7

- „County Reports“ (ca. 10.000)
- Scans von gebundenen Ausgaben

Dokumente

Table 3 Acreage, Population, Private Households and Dwellings

Local Authority Areas, Wards, Civil Parishes in Rural Districts, Conurbation Centres, New Towns

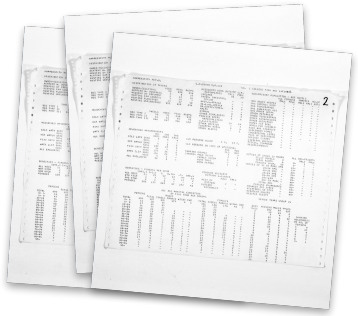
Notes:- (1) For definitions of area, dwellings, households and rooms, see pp. ix, x and xi.
(2) Households temporarily absent on Census night and their dwellings and rooms are included in cols. h, k and l but not in cols. j, m and n.
(3) Changes since 1951 are indicated by symbols - * boundary, † name, ‡ newly constituted area. Particulars of all these changes (except those relating to wards) are given in Table 4.

Area	Acreage	Population 1951 Persons	Population 1961 Persons	Population 1961 Males	Population 1961 Females	Population 1961 Persons per acre	Private households and dwellings, 1961: Private households	Population in private households	Structurally separate dwellings occupied	Rooms occupied	Density of occupation: Persons per room	Density of occupation: Percentage of persons at more than 1½ per room
a	b	c	d	e	f	g	h	j	k	l	m	n
***BEDFORDSHIRE**	**302,940**	**311,937**	**380,837**	**190,549**	**190,288**	**1·3**	**120,042**	**369,243**	**117,433**	**561,582**	**0·67**	**5·1**
***M.Bs. and U.Ds.**	**30,651**	**212,287**	**257,362**	**128,069**	**129,293**	**8·4**	**81,525**	**251,146**	**78,975**	**380,766**	**0·67**	**5·0**
***Rural Districts**	**272,290**	**99,650**	**123,475**	**62,480**	**60,995**	**0·5**	**38,517**	**118,097**	**38,458**	**180,816**	**0·66**	**5·3**
M.Bs. and U.Ds.												
Ampthill U.D.	**1,904**	**2,873**	**3,852**	**1,897**	**1,955**	**2·0**	**1,258**	**3,693**	**1,243**	**6,181**	**0·61**	**1·5**
Bedford M.B.	**4,972**	**53,075**	**63,334**	**31,089**	**32,245**	**12·7**	**20,391**	**60,952**	**18,617**	**94,426**	**0·66**	**8·1**
Wards:												
No. 1 Castle	215	6,581	5,853	2,948	2,905	27·2	2,102	5,240	1,511	7,131	0·75	26·5
No. 2 Newnham	706	8,296	8,408	3,922	4,486	11·9	2,883	8,245	2,732	15,014	0·56	2·8
No. 3 Kingsbrook	1,110	9,809	9,792	4,911	4,881	8·8	2,812	9,597	2,779	12,779	0·76	5·8
No. 4 Cauldwell	510	9,752	9,900	4,915	4,985	19·4	2,988	9,576	2,848	14,582	0·66	8·2
No. 5 Queens Park	527	5,849	6,648	3,251	3,397	12·6	2,269	6,588	2,136	11,430	0·58	5·3
No. 6 Harpur	542	6,131	6,579	3,225	3,354	12·1	2,291	6,141	1,856	9,972	0·63	15·0
No. 7 De Parys	1,362	6,657	16,154	7,917	8,237	11·9	5,046	15,565	4,755	23,518	0·68	4·7
Biggleswade U.D.	**4,759**	**7,431**	**8,050**	**3,957**	**4,093**	**1·7**	**2,564**	**7,756**	**2,560**	**12,149**	**0·64**	**2·9**
***Dunstable M.B.**	**2,092**	**17,190**	**25,645**	**12,796**	**12,849**	**12·3**	**8,024**	**25,407**	**7,944**	**37,756**	**0·68**	**3·0**
Wards:												
*No. 1 Northfields	270	2,737	3,672	1,848	1,824	13·6	1,098	3,643	1,093	5,080	0·72	4·1
No. 2 Chiltern	277	3,632	4,541	2,221	2,320	16·4	1,396	4,503	1,388	6,576	0·69	4·4
No. 3 Park	150	2,515	1,938	968	970	12·9	675	1,863	652	3,226	0·59	1·9
No. 4 Icknield	627	2,734	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0·70	1·9
No. 5 Priory	318	2,821	4,930	2,431	2,499	15·5	1,568	4,890	1,560	7,410	[illegible]	2·1
No. 6 [illegible]	450	2,751	3,440	1,800	1,640	7·2	1,039	3,199	1,019	5,020	[illegible]	4·2
Kempston U.D.	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2,975	14,763	[illegible]	**1·1**
Leighton Buzzard U.D.	2,487	9,025	11,745	5,822	5,923	4·7	3,723	11,520	3,716	17,431	[illegible]	**3·8**
Luton M.B.	8,773	110,381	131,583	66,002	65,581	15·0	41,278	128,938	40,624	191,653	[illegible]	**4·6**
Wards:												
No. 1 Central	[illegible]	[illegible]	[illegible]	[illegible]	3,615	28·4	2,265	6,563	2,162	10,615	[illegible]	6·2
No. 2 [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	8,197	15·4	4,995	16,265	4,899	23,560	[illegible]	4·3
No. 3 Wardown	491	12,527	11,690	5,856	5,834	23·8	3,928	11,183	3,732	18,302	[illegible]	5·7
No. 4 High Town	293	9,204	7,807	3,895	3,912	26·6	2,690	7,662	2,663	12,845	[illegible]	4·5
No. 5 St. Mary's	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4·0
No. 6 South	636	10,715	10,207	5,039	5,168	16·0	3,352	9,978	3,249	15,481	0·65	4·9
No. 7 Leagrave	1,733	15,235	22,485	11,196	11,289	13·0	6,802	22,055	6,762	30,914	0·72	4·3
No. 8 Limbury	1,140	14,584	20,078	9,999	10,079	17·6	6,363	20,008	6,336	29,535	0·68	4·2
No. 9 Stopsley	1,861	15,518	24,182	12,186	11,996	13·0	7,213	24,160	7,189	33,662	0·72	4·8
Sandy U.D.	**4,362**	**3,667**	**3,963**	**1,999**	**1,964**	**0·9**	**1,297**	**3,862**	**1,296**	**6,407**	**0·61**	**3·4**

- Komplexeres Tabellenformat
- Niedrige Priorität

Ziel / Nutzen

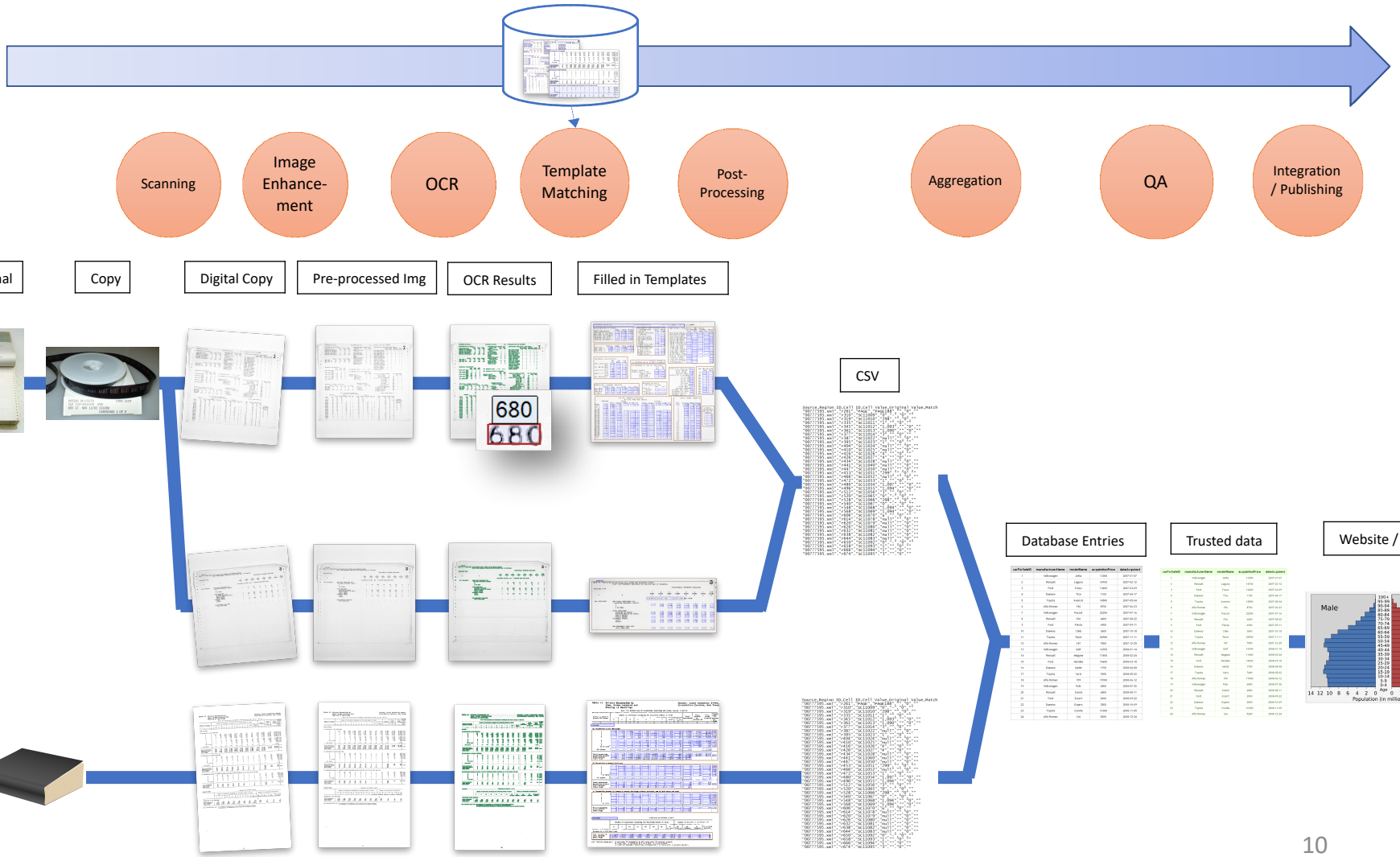
- Daten elektronisch zugänglich machen, wie neuere Zensus
- Datenbank
- Online, Suchfunktion, Export, …
- Öffentlich zugänglich (nach Projektende)
- Für Forschung oder sonstiges Interesse



Workflow

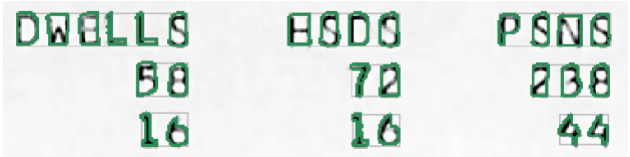
- Mehr als nur OCR
- Erkannter Text muss auf Tabellenmodelle abgebildet werden
- Bildqualität erlaubt keine 100%ige Automatisierung

Workflow

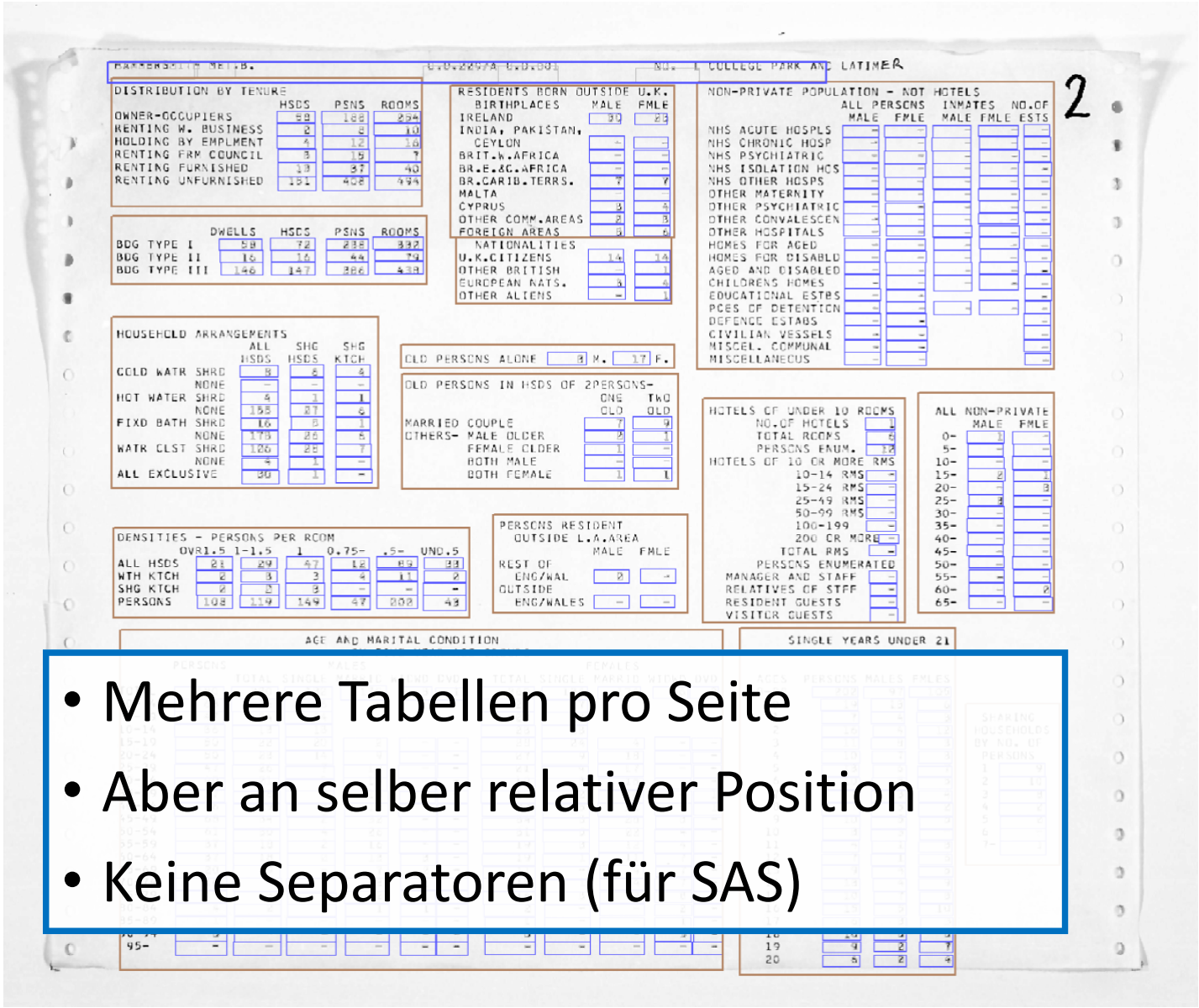

Scanning
Image Enhance-ment
OCR
Template Matching
Post-Processing
Aggregation
QA
Integration / Publishing
Original
Copy
Digital Copy
Pre-processed Img
OCR Results
Filled in Templates
680
680
CSV
Database Entries
Trusted data
Website / API
Male
Female
Population (in millions)

Vorverarbeitung und OCR

- Normierung auf TIFF Bilddateien
- Bildvorverarbeitung, um OCR zu verbessern
  - Verschiedene Verfahren getestet
  - Beste Kombination mit Ground Truth und OCR-Evaluierung ermittelt
- OCR
  - Tesseract und FineReader getestet (auch mit verschiedenen Einstellungen)
  - FineReader ist besser (und korrigiert Rotation)
  - Tesseract als sekundäres System in Nachverarbeitung
  - Wir brauchen: Text und Position von Glyphen
  - PAGE XML-Format

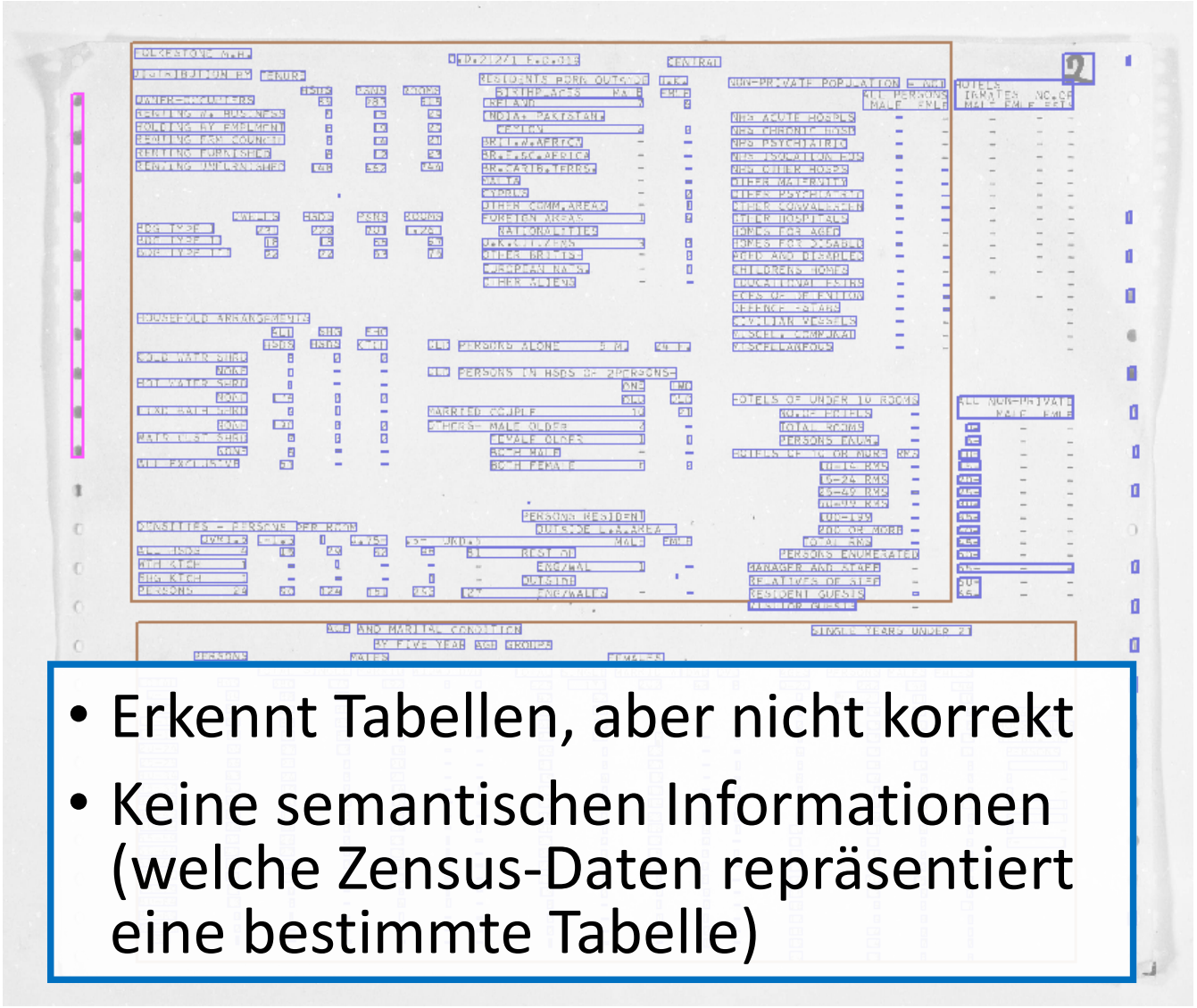


Tabellen



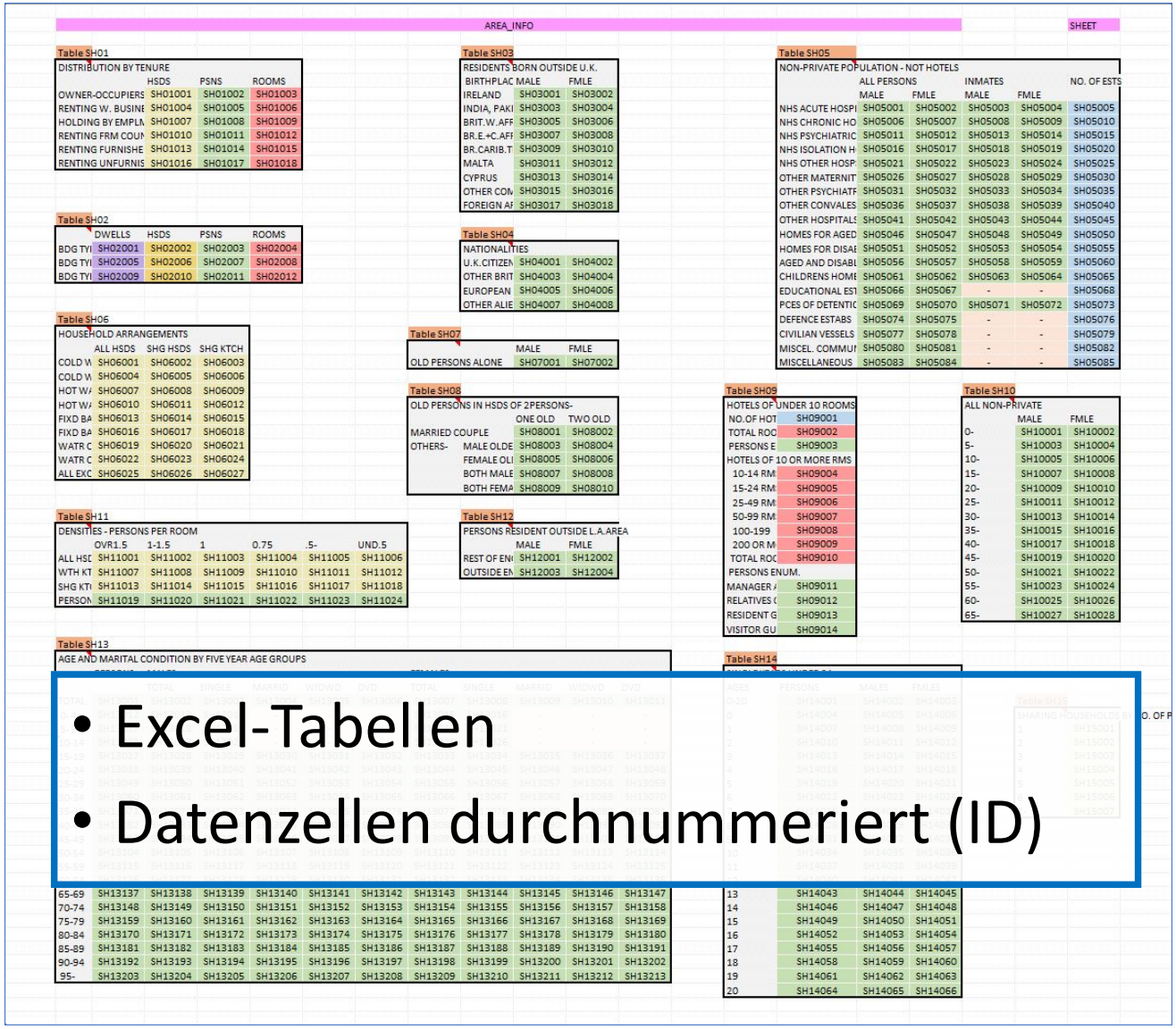
- Mehrere Tabellen pro Seite
- Aber an selber relativer Position
- Keine Separatoren (für SAS)

FineReader-Ergebnis

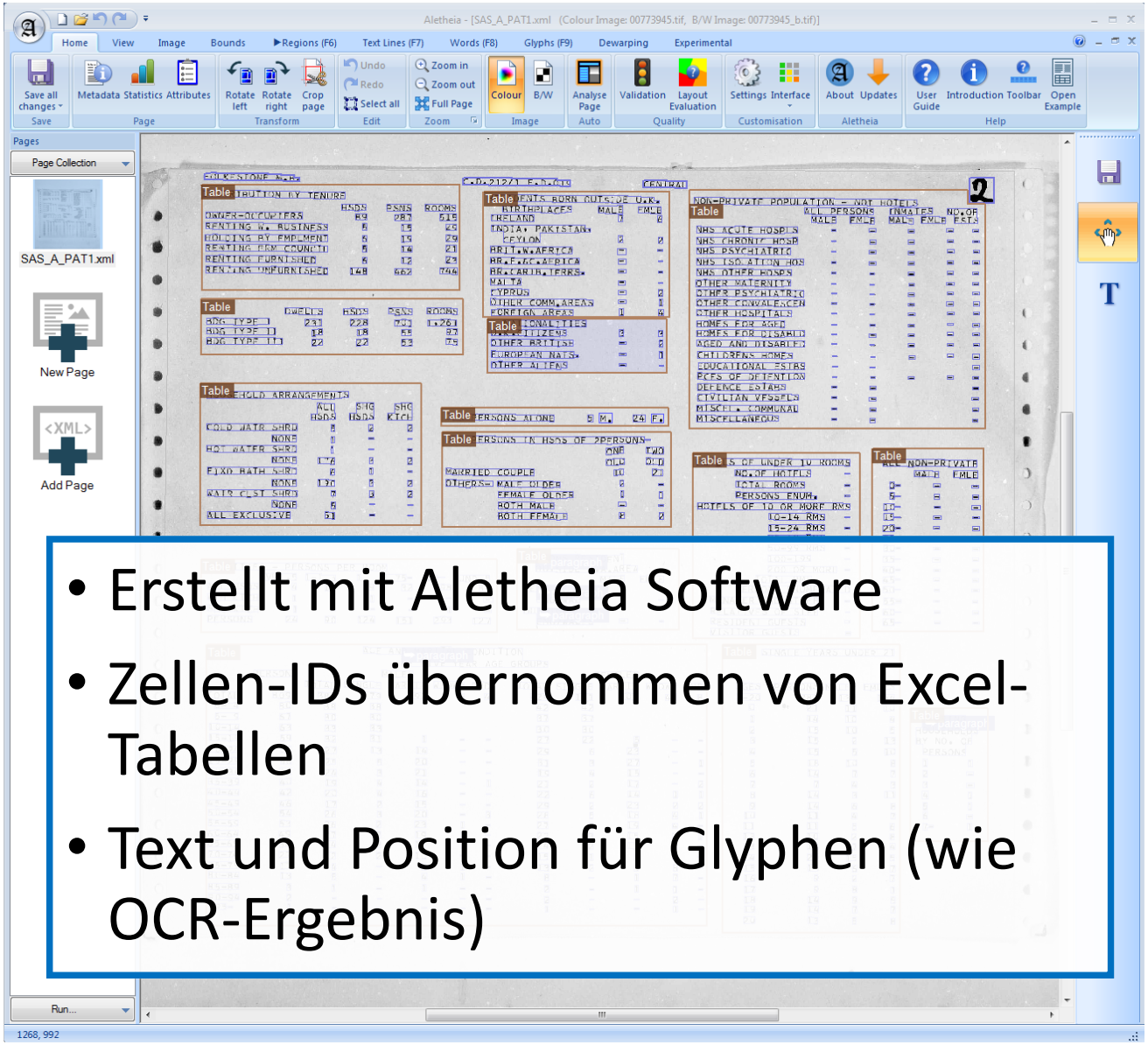


- Erkennt Tabellen, aber nicht korrekt
- Keine semantischen Informationen (welche Zensus-Daten repräsentiert eine bestimmte Tabelle)

Tabellenmodelle



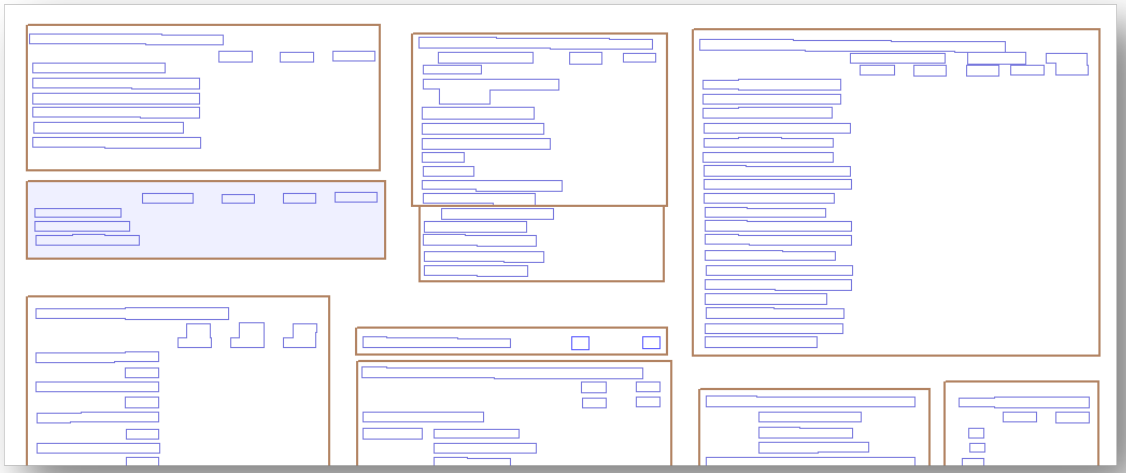
Templates

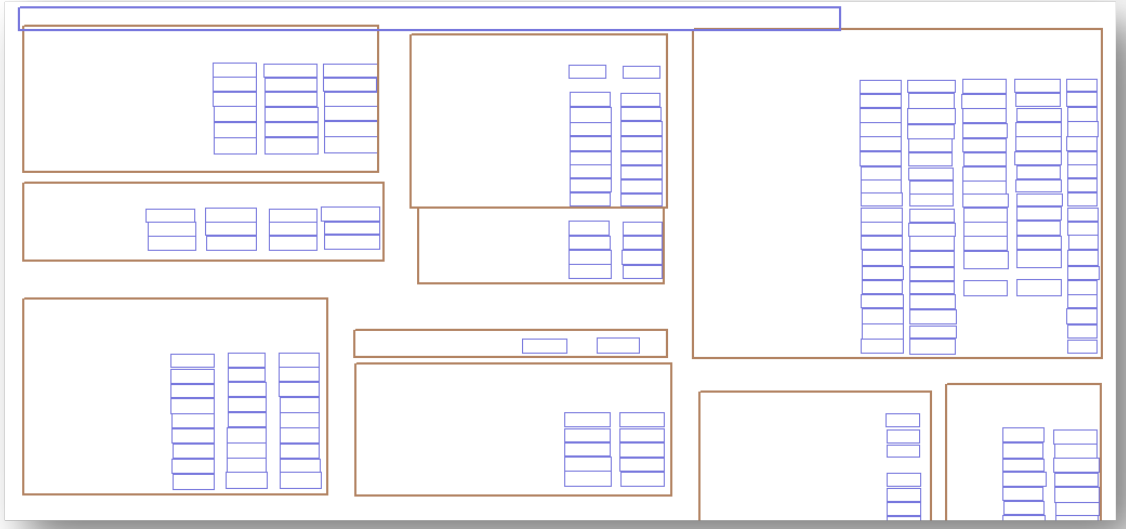


- Erstellt mit Aletheia Software
- Zellen-IDs übernommen von Excel-Tabellen
- Text und Position für Glyphen (wie OCR-Ergebnis)

Templates

Je zwei Templates: Numerische Zellen und textuelle Zellen

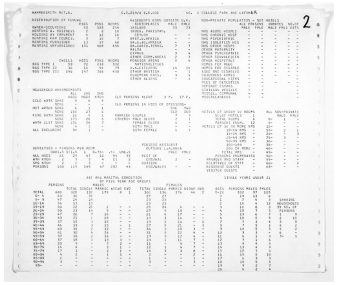


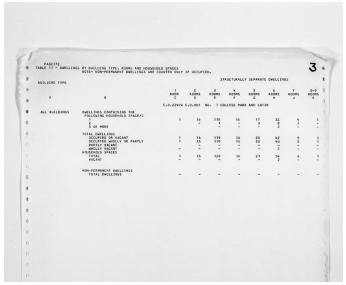


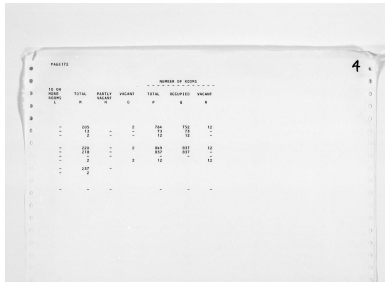
Wird wahrscheinlich in Zukunft zu einem Template vereinheitlicht.

Klassifizierung

- Metadaten für Bilddateien nicht ausreichend – Unbekannt welche Tabellen in Scan
  - Ist notwendig, um Templates auszuwählen
- Berechnung welche Tabellen am wahrscheinlichsten sind (für jede Seite)
  - Anhand vom unveränderlichen Tabelleninhalt (Zeilen- und Spaltenköpfe)
  - Vergleich OCR-Text mit Template-Text

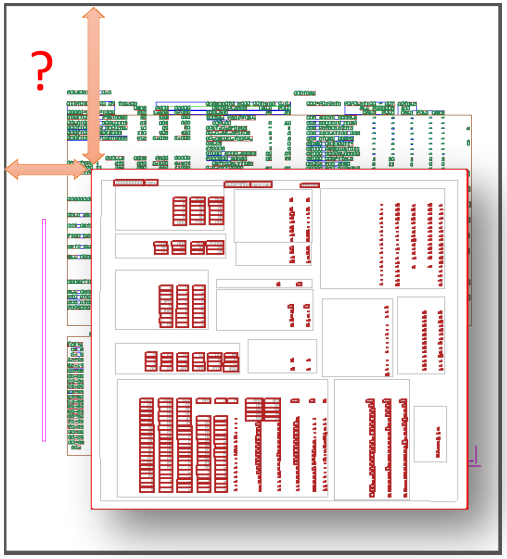


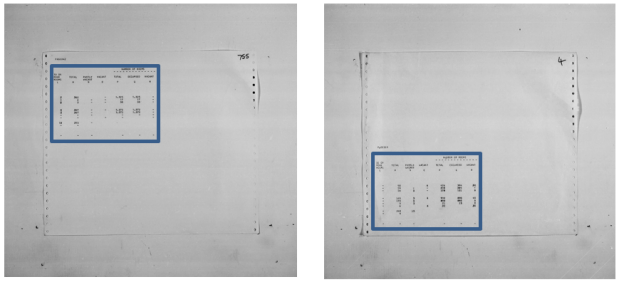




?

Template-Matching

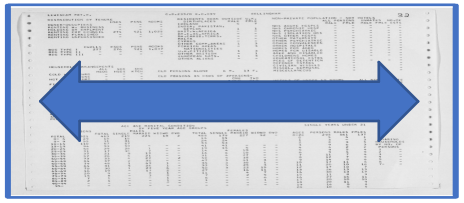


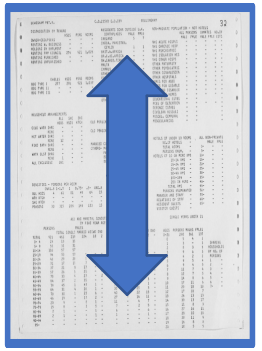


- Vergleich von OCR-Ergebnis und Template an verschiedenen Positionen
- Übereinstimmungswert anhand von Textinhalt
  - Glyphentext und –position…
- Bild wird nicht genutzt

Skalierung

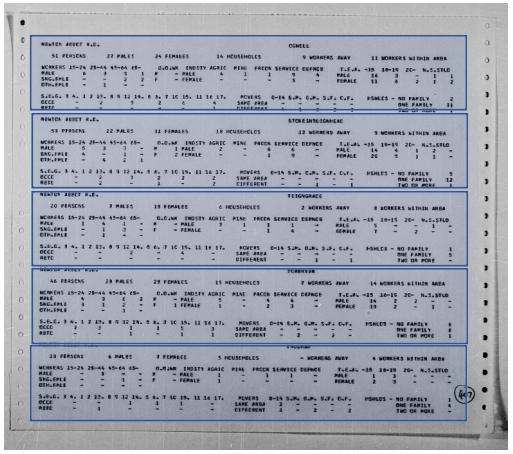
- Manche scans gestaucht oder gestreckt
- Bis zu 15%, in x- und/oder y-Richtung
- Tabellen-Templates passen nicht!
- Lösung: **Automatische Skalierung**
- Verschiedene Ansätze:
  - Mittels Abstand zwischen Wörtern, die nur einmal vorkommen
  - Mittels Durchschnitts-Breite von bestimmten Wörtern
  - Mittels Textrand-Position





Sonderfälle

- Wiederholung von Templates auf einer Seite
- Verschiedene Kombinationen von Templates auf einer Seite
- Lösung: Mehrfaches Matching mit kleineren Templates



Erfolgreich platziertes Template

Exclusion zone

Sonderfälle

- Unbekannte Zeilenanzahl
- Lösung: Zeilen-Template, dass Stück für Stück nach unten verschoben wird (bis Endbedingung erfüllt)

ABB-ABE

Name	Des-crip-tion	Adminis-trative County in which situated	Borough, Urban District or Rural District in which situated	No. of Regist-ration District in which situated	P.S.D. or P.S.B. code No.	Popula-tion 1961
Abbots Bickington	C.P.	Devon	Holsworthy R.D.	336/1	152	22
Abbots Bromley	C.P.	Staff.	Uttoxeter R.D.	446/3	609	1,071
Abbotsbury	C.P.	Dorset	Dorchester R.D.	296/1	174	470
Abbotsbury	Loc.	Devon	Newton Abbot U.D.	339/3	160	-
Abbots, Charlton	Loc.	Glouc.	Cheltenham R.D. (Sudeley C.P.)	347/1	258	-
Abbots, Duntisbourne	C.P.	Glouc.	Cirencester R.D.	348/1	241	211
Abbotsham	C.P.	Devon	Bideford R.D.	333/1	144	287
Abbots, Hanham	C.P.	Glouc.	Warmley R.D.	352/1	246	3,862
Abbots Haye	Loc.	Staff.	Cheadle R.D. (Cheadle C.P.)	446/1	597	-
Abbots, Hemingford	C.P.	Hunts.	St. Ives R.D.	174/2	320	628
Abbotside, High	C.P.	Yks.(N.R.)	Aysgarth R.D.	39/1	790	253
Abbotside, Low	C.P.	Yks.(N.R.)	Aysgarth R.D.	39/1	790	97
Abbotskerswell	Loc.	Devon	Newton Abbot R.D. (Kerswell C.P.)	339/3	160	-

Area	Acreage	Population					Private households and dwellings, 1961					
		1951	1961								Density of occupation	
		Persons	Persons	Males	Females	Persons per acre	Private house-holds	Popula-tion in private house-holds	Struc-turally separate dwellings occupied	Rooms occupied	Persons per room	Percent-age of persons at more than 1½ per room
a	b	c	d	e	f	g	h	i	k	l	m	n
BEDFORDSHIRE	302,140	511,531	560,057	280,549	280,288	1·5	120,043	369,443	117,433	581,582	0·97	5·1

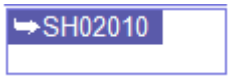
Text-Übertragung

- OCR-Ergebnis-Text wird in platziertes Template übertragen
- Jeweils in die nächstgelegene Tabellenzelle (Positionen der Glyphen bekannt)
- Semantische Information für den erkannten Text (anahnd der Zellen-ID)

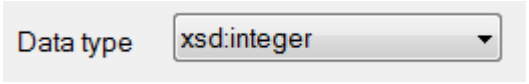
= Zelle T03_2755
= Anzahl männlicher Personen im Einzugsgebiet

Nachverarbeitung

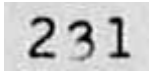
- Erwarteter Zelleninhalt bekannt
- Typisches Problem: Zahl erwartet, aber Buchstaben enthalten
- Verbesserungsansätze:
  - Ersetzungsregeln (z.B. großes i mit 1 ersetzen)
  - Tesseract OCR, pro Zelle, eingeschränkt auf Ziffern
  - Nutzen von Varianten im OCR-Ergebnis (falls verfügbar)



Zellen-ID

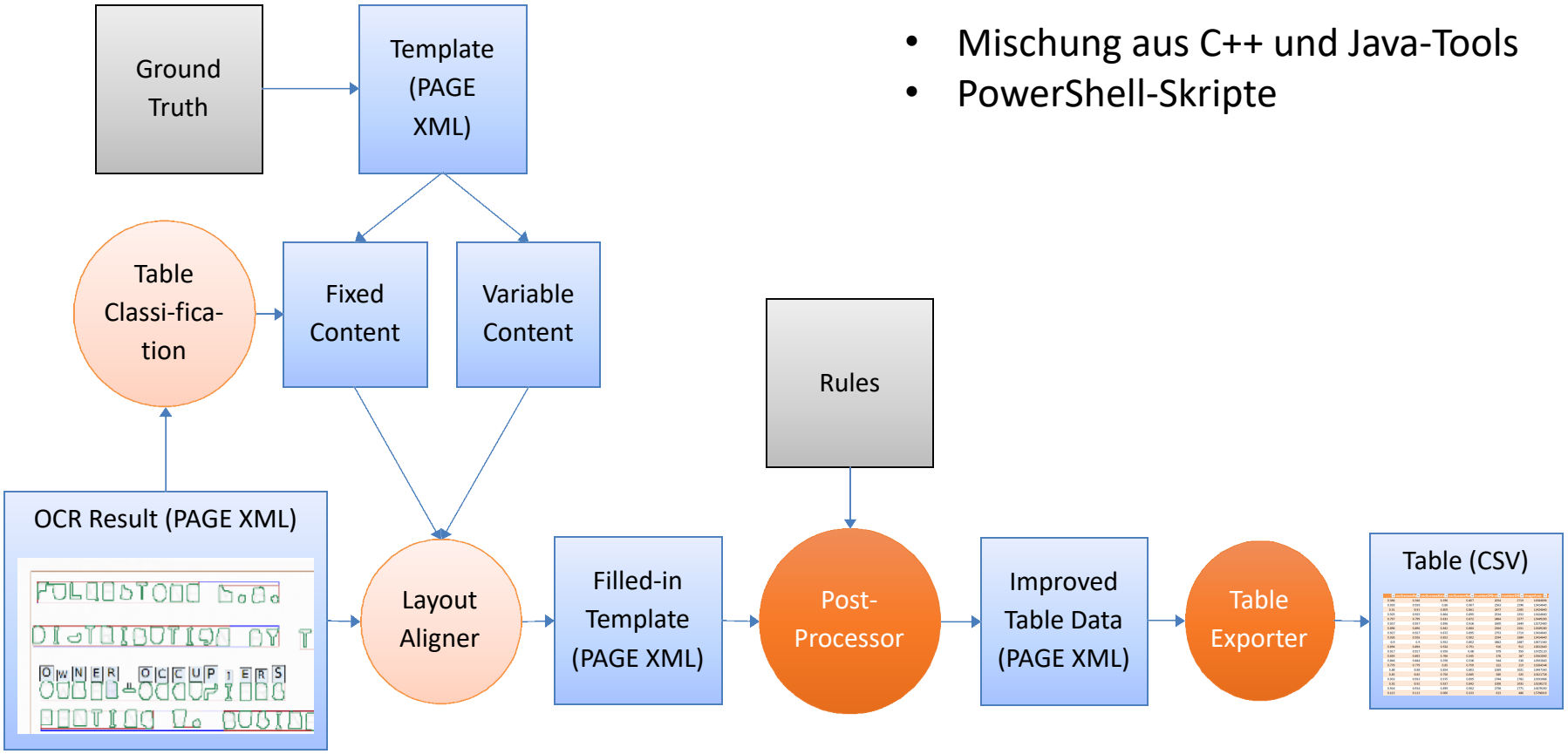


Datentyp



Bildausschnitt

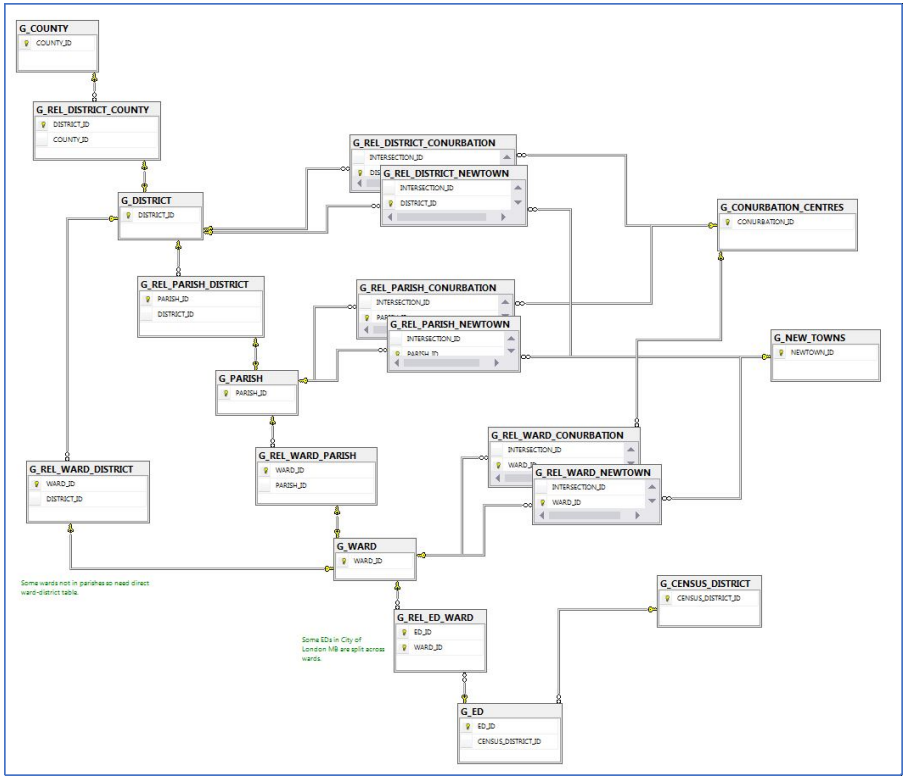
Übersicht der “Pipeline”



- Mischung aus C++ und Java-Tools
- PowerShell-Skripte

Datenbank

- Import der extrahierten Daten
- Zensus-Modell



- Bisher Microsoft SQL Server
- Im Moment Umstieg auf MySQL

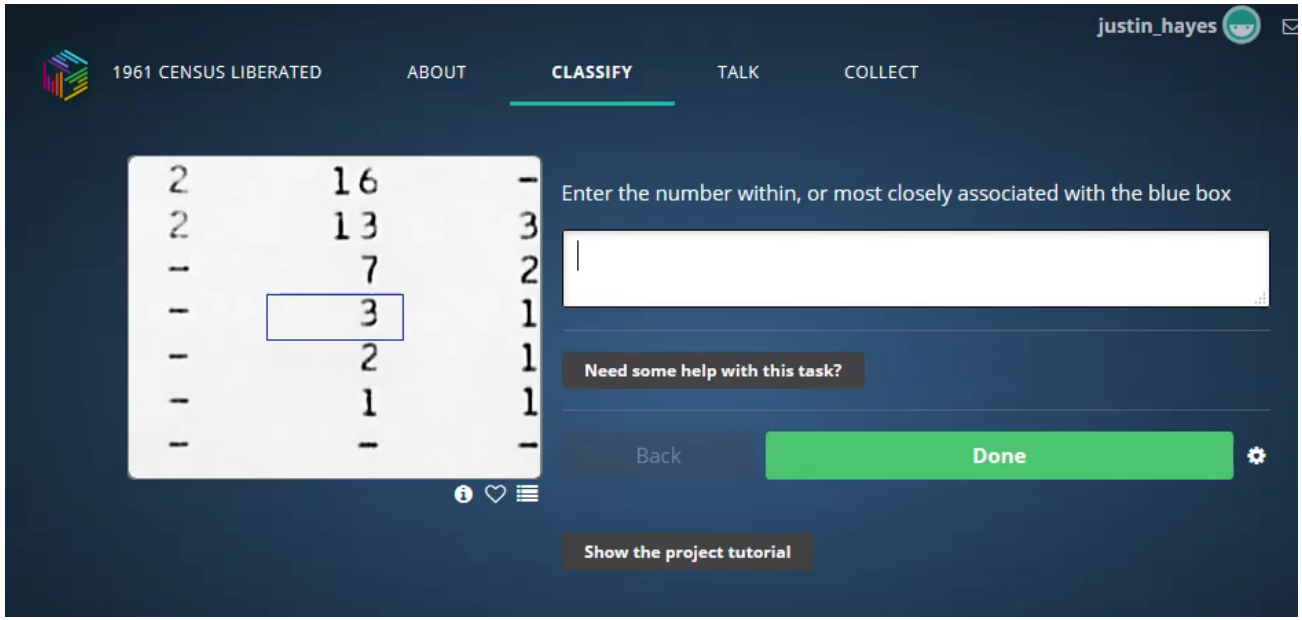
Qualitätskontrolle

- Korrektheit der Zellen etwa 98%
- 2% falsche Zellen von Millionen Zellen ist viel!
- Confidence-Werte in der Pipeline (aber nur Anhaltspunkte)
- Besser: **Automatische Kontrolle**, um falsche Zellen zu finden
  - Nutzen von Datenredundanzen, um Fehler zu finden
  - Summen innerhalb einer Tabelle oder Summen von Tabellen für Gebiete unterschiedlicher Größe

$$\sum \frac{\left(T/S\right)^2 - 1}{V}$$

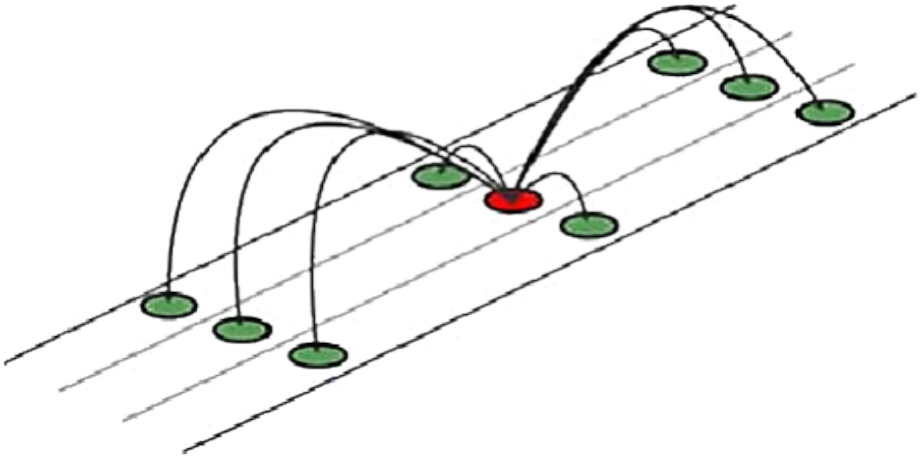
Nachkorrektur

- Direkt oder mittels Crowd-Sourcing
- Zooniverse-Plattform
- Zelle für Zelle (mit Bildausschnitten vom originalen Scan)



Lücken Füllen

- Automatische Berechnung von fehlenden Werten
- (Sobald genug Daten korrigiert wurden)



Status

- Kompletter Arbeitsablauf implementiert
- Mit gesponsort von ABBYY
- Bereit für Anschlussprojekt (ca. Jahr)
  - Alle Scans verarbeiten
  - Verfeinerung der Tools und des Workflows
  - OCR Training? (Tesseract vielversprechend)
- Zensus-Daten werden dann öffentlich gemacht

Fragen?

www.primaresearch.org