

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...)

UK Data Service



PRImA
Research Lab



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 EMPMLNT	4	12	16
RENTING FRM COUNCIL	3	15	7
RENTING FURNISHED	13	37	40
RENTING UNFURNISHED	151	408	494

	BIRTHPLACES	OUTSIDE U.K.
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	INMATES		
MALE	FMLE	MALE	FMLE	ESTS
NHS ACUTE HOSPLS	-	-	-	-
NHS CHRONIC HOSP	-	-	-	-
NHS PSYCHIATRIC	-	-	-	-
NHS ISOLATION HCS	-	-	-	-
NHS OTHER HCSPS	-	-	-	-
OTHER MATERNITY	-	-	-	-
OTHER PSYCHIATRIC	-	-	-	-
OTHER CONVALESCEN	-	-	-	-
OTHER HOSPITALS	-	-	-	-
HOMES FOR AGED	-	-	-	-
HOMES FOR DISABLED	-	-	-	-
AGED AND DISABLED	-	-	-	-
CHILDREN'S HOMES	-	-	-	-
EDUCATIONAL ESTBS	-	-	-	-
PCES OF DETENTION	-	-	-	-
DEFENCE ESTABS	-	-	-	-
CIVILIAN VESSELS	-	-	-	-
MISCEL. COMMUNAL	-	-	-	-
MISCELLANEOUS	-	-	-	-

HOUSEHOLD ARRANGEMENTS

	ALL	SHG	SHG
	HSDS	HSDS	KTCH
COLD WTR 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
WTR CLST SHRD	126	28	7
NONE	4	1	-
ALL EXCLUSIVE	30	1	-

	CLD PERSONS	ALONE	3 M.	17 F.
	HSDS	HSDS	KTCH	
CLD PERSONS IN HSDS OF 2PERSONS-				
CNE ONE TWO				
CLD OLD				

	HOTELS OF UNDER 10 ROOMS		ALL NON-PRIVATE	
	NO. OF HOTELS	1	MALE	FMLE
TOTAL ROOMS	6		0-	1-
PERSONS ENUM.	12		5-	-
	HOTELS OF 10 OR MORE RMS			
10-14 RMS	-	15-	2	1
15-24 RMS	-	20-	-	3
25-49 RMS	-	25-	3	-
50-99 RMS	-	30-	-	-
100-199	-	35-	-	-
200 OR MORE	-	40-	-	-
TOTAL RMS	-	45-	-	-
PERSONS ENUMERATED	50-	-	-	-
MANAGER AND STAFF	-	55-	-	-
RELATIVES OF STFF	-	60-	-	2
RESIDENT GUESTS	-	65-	-	-
VISITER GUESTS	-	-	-	-

DENSITIES - PERSONS PER RCDM

OVER 1.5	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

SINGLE YEARS UNDER 21

PERSONS

MALES

FEMALES

	TOTAL	SINGLE	MARRIED	WIDOW	DVD		TOTAL	SINGLE	MARRIED	WIDOW	DVD	AGES	PERSONS	MALES	FMLES
TOTAL	680	320	132	179	8	1	360	136	176	46	2	0-20	202	97	105
0- 4	63	36	36	-	-	-	27	27	-	-	-	0	19	13	6
5- 9	47	24	24	-	-	-	23	23	-	-	-	1	7	4	3
10-14	36	13	13	-	-	-	23	23	-	-	-	2	16	4	12
15-19	50	22	20	2	-	-	28	24	4	-	-	3	11	8	3
20-24	50	23	14	9	-	-	27	9	18	-	-	4	10	7	3
25-29	47	26	7	19	-	-	21	4	17	-	-	5	13	6	7
30-34	40	21	2	19	-	-	19	3	16	-	-	6	10	5	2
35-39	37	18	3	15	-	-	19	2	16	-	-	1	7	3	8
40-44	46	19	3	15	-	-	27	5	21	-	-	1	8	7	5
45-49	68	34	2	32	-	-	34	3	28	3	-	9	10	5	5
50-54	61	30	4	26	-	-	31	5	22	4	-	10	3	3	-
55-59	37	19	2	15	-	-	19	2	12	6	-	11	5	3	-
60-64	37	18	2	13	3	-	19	1	11	7	-	12	7	1	6
65-69	20	9	2	2	-	-	1	1	7	7	-	13	9	4	-
70-74	-	-	-	-	-	-	-	-	-	-	-	5	5	3	-
75-79	1	-	-	-	-	-	1	-	-	-	-	17	6	3	1
80-84	3	-	-	-	-	-	3	-	-	-	-	10	5	5	-
85-94	3	-	-	-	-	-	3	-	-	-	-	19	9	2	7
95-	-	-	-	-	-	-	-	-	-	-	-	20	6	2	4

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

- 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	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					NUMBER OF ROOMS		
	TOTAL L	PARTLY VACANT M	VACANT N	TOTAL O	OCCUPIED P	VACANT Q	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*	Total
		a	b	c	d	e	f	g	h	j	i
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	590
	2 " "	78	192	959	2,498	1,707	756	641	135	4	4
	3 " "		79	149	266	348	246	333	166	166	166
Wholly residential permanent buildings containing one dwelling	4 or more " "			114	524	468	491	670	590	590	590
	Total Dwellings (occupied or vacant)	3,950	15,830	57,943	145,745	269,203	82,266	19,755	11,225	3,557	6
	(occupied (wholly or partly))	3,900	15,435	56,335	143,316	266,435	81,137	19,485	11,030	3,507	6
	(partly vacant)		3	8	29	53	98	72	120	89	89
Not wholly residential permanent buildings containing one dwelling	(wholly vacant)	50	395	1,608	2,429	2,768	1,129	270	195	50	50
	Household Spaces (total)	3,950	15,908	58,293	147,344	273,346	86,502	23,066	15,532	7,628	6
	(vacant)	50	398	1,616	2,465	2,833	1,239	357	335	177	6
	Dwellings containing										
All Buildings	1 Household space(s)	1,108	7,468	42,079	127,446	257,737	77,089	17,306	8,960	2,415	5
	2 " "	23	111	847	2,428	1,666	738	621	128	128	128
	3 " "		65	128	239	331	240	324	160	160	160
	4 or more " "			103	513	451	477	651	570	570	570
Wholly residential permanent buildings containing one dwelling	Total Dwellings (occupied or vacant)	1,108	7,491	42,255	128,524	260,717	79,537	18,759	10,556	3,273	5
	(occupied (wholly or partly))	1,090	7,336	41,256	126,566	258,083	78,448	18,496	10,368	3,225	5
	(partly vacant)		1	4	26	49	90	70	114	87	87
	(wholly vacant)	18	155	999	1,558	2,634	1,089	263	188	48	48
Not wholly residential permanent buildings containing one dwelling	Household Spaces (total)	1,108	7,514	42,496	129,936	264,696	83,633	21,986	14,751	7,212	5
	(vacant)	18	156	1,003	1,990	2,693	1,191	347	322	171	5
	Dwellings containing										
	1 Household space(s)	112	882	3,625	6,448	4,955	2,010	699	459	165	18
All Buildings	2 " "	15	8	36	36	36	32	14	14	7	7
	3 " "		6	10	10	5	2	2	2	2	2
	4 or more " "			4	7	8	2	4	4	5	5
	Total Dwellings (occupied or vacant)	112	887	3,631	6,464	4,969	2,029	711	466	179	18
Wholly residential permanent buildings containing one dwelling	Dwellings containing										
	1 Household space(s)	105	851	3,481	6,229	4,909	2,029	711	466	179	18
	2 " "	1	4	1	2	1	2	1	1	1	1
	3 " "		1	2	2	1	2	1	1	-	-
Not wholly residential permanent buildings containing one dwelling	(occupied (wholly or partly))	105	851	3,481	6,229	4,909	2,126	743	500	216	19
	(partly vacant)	1	4	1	2	1	2	1	1	1	1
	(wholly vacant)	25	194	451	202	35	14	1	3	2	2
	Household Spaces (total)	1,262	6,667	11,781	10,794	3,552	741	337	270	200	35
	(vacant)	25	196	454	204	38	18	1	6	4	35

- „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					Private households and dwellings, 1961					
		1951		1961			Private households	Population in private households	Structurally separate dwellings occupied	Rooms occupied	Density of occupation	
		Persons	Persons	Males	Females	Persons per acre					Persons per room	Percent of persons at more than 1½ per room
a	b	c	d	e	f	g	h	i	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.B.s. and U.Ds.	30,651	212,267	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.B.s. and U.Ds.												
Amphill 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.	4,872	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	4,466	11·9	2,683	6,245	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	5,649	6,648	3,251	3,597	12·6	2,269	6,588	2,136	11,430	0·58	5·3
No. 6 Harpur	542	6,151	6,579	3,225	3,554	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	6,237	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	2,520	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	5,226	0·59	1·9
No. 4 Tolkafield	627	2,724	—	—	—	—	2,269	2,700	2,610	12,444	0·70	1·9
King's Sutton U.D.	316	2,821	4,930	2,351	2,440	15·3	1,588	4,800	1,860	7,410	0·77	2·1
Leighton Buzzard U.D.	3,487	9,025	11,745	5,822	5,923	4·7	3,723	11,520	3,716	17,431	0·57	3·8
Luton M.B.	8,773	110,381	131,583	66,902	65,581	15·0	41,278	128,938	40,624	191,653	0·58	4·6
Wards:												
No. 1 Great Linford	1,077	3,615	28·4	2,265	2,563	2,162	10,615	0·52	6·2			
No. 2 Leagrave	891	8,197	15·4	4,995	16,285	4,899	18,580	0·70	4·3			
No. 3 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. 4 High Town	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

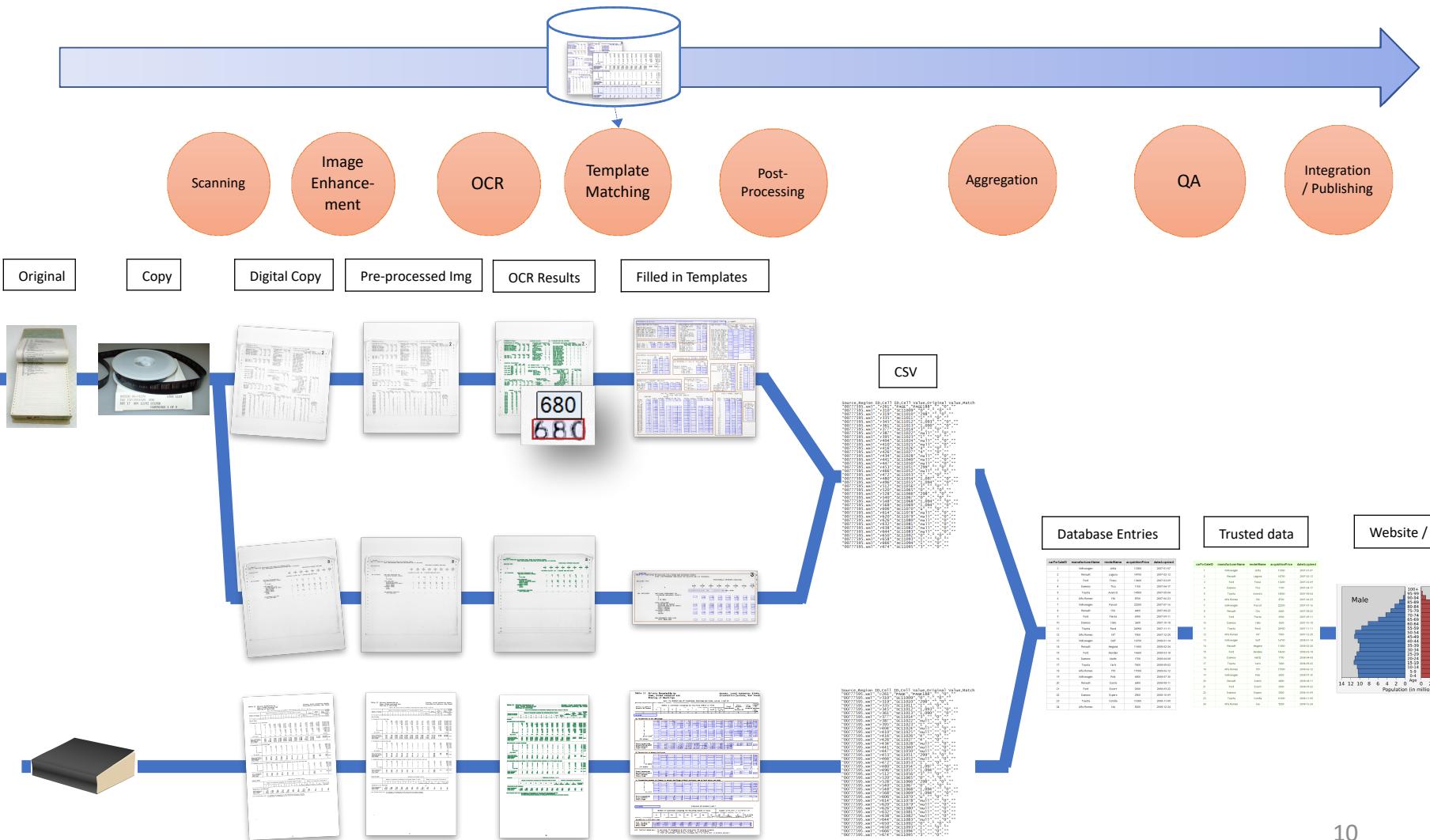


Workflow



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

Workflow



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

PATTERSON MEI.B.				C.0.2467A C.0.001				NO. 1 COLLEGE PARK AND LATIMER			
DISTRIBUTION BY TENURE				RESIDENTS BORN OUTSIDE U.K.				NON-PRIVATE POPULATION - NOT HOTELS			
OWNER-OCCUPIERS	HSDS	PSNS	ROOMS	BIRTHPLACES	MALE	FMLE	ALL PERSONS	INMATES	NO. OF	MALE	FMLE ESTS
KENTING W. BUSINESS	2	2	10	IRELAND	39	23	MALE	FMLE	NO. OF	MALE	FMLE ESTS
HOLDING BY EMPLOYL	6	6	16	INDIA, PAKISTAN,	-	-	2	1	1	0	0
RENTING FRM COUNCIL	3	3	15	CEYLON	-	-	3	2	1	0	0
RENTING FURNISHED	13	37	40	BRIT.W.AFRICA	-	-	4	3	1	0	0
RENTING UNFURNISHED	151	468	494	BRA&G.AFRICA	-	-	5	4	1	0	0
DWELLS	HSDS	PSNS	ROOMS	BR&CARIB.TERRS.	7	7	6	5	1	0	0
BDG TYPE I	59	72	232	MALTA	-	-	7	6	1	0	0
BDG TYPE II	15	15	79	CYPRUS	-	-	8	7	1	0	0
BDG TYPE III	146	147	439	OTHER COMM. AREAS	2	2	9	8	1	0	0
FOREIGN AREAS				NATIONALITIES	14	14	10	9	1	0	0
U.K.CITIZENS				OTHER BRITISH	-	-	11	10	1	0	0
EUROPEAN RATS.				OTHER ALIENS	-	-	12	11	1	0	0
HOUSEHOLD ARRANGEMENTS				OLD PERSONS ALONE				NON-PRIVATE POPULATION - HOTELS			
ALL	HSDS	SHG	SHG	ONE	8	17	F	NO. OF HOTELS	1	0	0
COLD WTR SHRD	8	8	8	TWO	7	9		TOTAL ROOMS	4	0	0
NONE	-	-	-	OLD	2	1		PERSONS ENUM.	10	0	0
HOT WATER SHRD	4	3	1	OLD	1	-		HOTELS OF 10 OR MORE RMS	-	0	0
NONE	155	87	56	OLD	-	-		10-14 RMS	-	0	0
FIXED BATH SHRD	16	5	1	OLD	-	-		15-24 RMS	-	0	0
NONE	173	25	5	MARRIED COUPLE	7	9		25-49 RMS	-	0	0
WTR CLST SHRD	126	28	7	OTHERS- MALE OLDER	2	1		50-99 RMS	-	0	0
NONE	4	1	-	MALE OLDER	1	-		100-199	-	0	0
ALL EXCLUSIVE	20	1	-	BOTH MALE	-	-		200 OR MORE	-	0	0
				BOTH FEMALE	1	1		TOTAL RMS	-	0	0
DENSITIES - PERSONS PER ROOM				PERSONS RESIDENT OUTSIDE L.A. AREA				PERSONS ENUMERATED	-	0	0
OVER 1.5	1-1.5	1	0.75-	REST OF ENGL/WAL	MALE	FMLE		MANAGER AND STAFF	-	0	0
ALL HSDS	21	29	47	OUTSIDE	0	-		RELATIVES OF STAFF	-	0	0
WTH KITCH	2	3	3	ENGL/WALES	0	-		RESIDENT GUESTS	-	0	0
SHG KITCH	2	2	3					VISITOR GUESTS	-	0	0
PERSONS	108	119	149							0	0
AGE AND MARITAL CONDITION				SINGLE YEARS UNDER 21				SHARING HOUSEHOLDS BY NO. OF PERSONS			
10-14	-	-	-	10	20	30	40	50	60	70	80
15-19	-	-	-	11	21	31	41	51	61	71	81
20-24	-	-	-	12	22	32	42	52	62	72	82
25-29	-	-	-	13	23	33	43	53	63	73	83
30-34	-	-	-	14	24	34	44	54	64	74	84
35-39	-	-	-	15	25	35	45	55	65	75	85
40-44	-	-	-	16	26	36	46	56	66	76	86
45-49	-	-	-	17	27	37	47	57	67	77	87
50-54	-	-	-	18	28	38	48	58	68	78	88
55-59	-	-	-	19	29	39	49	59	69	79	89
60-64	-	-	-	20	30	40	50	60	70	80	90
65-69	-	-	-								
70-74	-	-	-								
75-79	-	-	-								
80-84	-	-	-								
85-89	-	-	-								
90-94	-	-	-								
95-99	-	-	-								

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

FineReader-Ergebnis

DECEASED, M.R.				ED 21271 F 0003				CENTRAL			
DEATHS BY CAUSE				BIRTHS ABTS MALE				NON-PRIVATE POPULATION			
DEATHS BY AGE				IRISH				K/L PERSONS			
DEATHS BY SEX				INDIA PAKISTAN				MALE FEMALE			
DEATHS BY ETHNICITY				IRELAND				HOTELS			
DEATHS BY RELIGION				CEYLON				IRATES NO OF			
DEATHS BY COUNTRY				BRIT & AFTRICA				MALE FEMALE			
DEATHS BY TERRITORIES				BR LAT AMERICA				HOTELS			
DEATHS BY ISLAMIC SHIPS				LATVIA LIBERIA				IRATES NO OF			
DEATHS BY AGE				MALT				MALE FEMALE			
DEATHS BY GENDER				CYPRUS				HOTELS			
DEATHS BY RACE				OTHER COMM AREA				IRATES NO OF			
DEATHS BY ETHNIC GROUP				FOREIGN REGNS				MALE FEMALE			
DEATHS BY RELIGION				RAJAHOMALI				HOTELS			
DEATHS BY CITIZENSHIP				PEK CHIN TWS				IRATES NO OF			
DEATHS BY BRITISH				OLIVE BRITISH				MALE FEMALE			
DEATHS BY NON-BRITISH				CZECHOSLOVAKIA				HOTELS			
DEATHS BY NATIONALITY				GER ALIENS				IRATES NO OF			
DEATHS BY ARRANGEMENTS				PERSONS ALONE				MALE FEMALE			
DEATHS BY MARITAL STATUS				PERSONS IN HDS OR 2 PERSONS				HOTELS OF UNDER 10 ROOMS			
DEATHS BY SEX				PARTNED COUPLE				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY AGE GROUP				2 HRS - MALE OLDER				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				FEMALE OLDER				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				SC H H WLD				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				SC H FEMAL				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				PERSONS RESIDENT				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				OUTSIDE LGA AREA				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				MALE FEMAL				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				REST OF ENGLAND				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				WELSH				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				SCOTLAND				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				ENG WALES				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				SCOT WALES				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				PERSONS RESIDENT				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				OUTSIDE LGA AREA				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				MALE FEMAL				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				REST OF ENGLAND				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				WELSH				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				SCOT WALES				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				PERSONS RESIDENT				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				OUTSIDE LGA AREA				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				MALE FEMAL				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				REST OF ENGLAND				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				WELSH				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				SCOT WALES				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				PERSONS RESIDENT				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				OUTSIDE LGA AREA				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				MALE FEMAL				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				REST OF ENGLAND				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				WELSH				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				SCOT WALES				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				PERSONS RESIDENT				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				OUTSIDE LGA AREA				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				MALE FEMAL				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				REST OF ENGLAND				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				WELSH				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				SCOT WALES				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				PERSONS RESIDENT				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				OUTSIDE LGA AREA				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				MALE FEMAL				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				REST OF ENGLAND				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				WELSH				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				SCOT WALES				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				PERSONS RESIDENT				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				OUTSIDE LGA AREA				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				MALE FEMAL				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				REST OF ENGLAND				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				WELSH				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				SCOT WALES				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				PERSONS RESIDENT				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				OUTSIDE LGA AREA				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				MALE FEMAL				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				REST OF ENGLAND				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				WELSH				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				SCOT WALES				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				PERSONS RESIDENT				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				OUTSIDE LGA AREA				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				MALE FEMAL				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				REST OF ENGLAND				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				WELSH				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				SCOT WALES				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY NATIONALITY				PERSONS RESIDENT				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY RELATIONSHIP				OUTSIDE LGA AREA				HOTELS OF 10 OR MORE ROOMS			
DEATHS BY GENDER				MALE FEMAL							

Tabellenmodelle

AREA_INFO										SHEET									
Table SH01										Table SH03									
DISTRIBUTION BY TENURE										RESIDENTS BORN OUTSIDE U.K.									
HEDS	PSNS	ROOMS								BIRTHPLAC	MALE	FMLE							
OWNER-OCCUPIERS	SH01001	SH01002	SH01003							IRELAND	SH03001	SH03002							
RENTING W. BUSINE	SH01004	SH01005	SH01006							INDIA, PAKI	SH03005	SH03004							
HOLDING BY EMPLN	SH01007	SH01008	SH01009							BRIT.W. AFF	SH03005	SH03006							
RENTING FRM COUN	SH01010	SH01011	SH01012							BR.E + CAFF	SH03007	SH03008							
RENTING FURNISHE	SH01013	SH01014	SH01015							BR.CARIB.T	SH03009	SH03010							
RENTING UNFURNIS	SH01016	SH01017	SH01018							MALTA	SH03011	SH03012							
										CYPRUS	SH03013	SH03014							
										OTHER COM	SH03015	SH03016							
										FOREIGN AF	SH03017	SH03018							
Table SH02										DWELLS	HSDS	PSNS	ROOMS						
BDG TYI	SH02001	SH02002	SH02003	SH02004															
										BDG TYI	SH02005	SH02006	SH02007	SH02008					
										BDG TYI	SH02009	SH02010	SH02011	SH02012					
Table SH05										Table SH06									
NON-PRIVATE POPULATION - NOT HOTELS										ALL PERSONS		INMATES		NO. OF ESTS					
MALE										MALE		MALE		NO. OF ESTS					
NHS ACUTE HOSPI	SH05001	SH05002	SH05003	SH05004	SH05005														
NHS CHRONIC HO	SH05006	SH05007	SH05008	SH05009	SH05010														
NHS PSYCHIATRIC	SH05011	SH05012	SH05013	SH05014	SH05015														
NHS ISOLATION H	SH05016	SH05017	SH05018	SH05019	SH05020														
NHS OTHER HOSP	SH05021	SH05022	SH05023	SH05024	SH05025														
OTHER MATERNIT	SH05026	SH05027	SH05028	SH05029	SH05030														
OTHER PSYCHIATF	SH05031	SH05032	SH05033	SH05034	SH05035														
OTHER CONVALES	SH05036	SH05037	SH05038	SH05039	SH05040														
OTHER HOSPITALS	SH05041	SH05042	SH05043	SH05044	SH05045														
HOMES FOR AGED	SH05046	SH05047	SH05048	SH05049	SH05050														
HOMES FOR DISAB	SH05051	SH05052	SH05053	SH05054	SH05055														
AGED AND DISABL	SH05056	SH05057	SH05058	SH05059	SH05060														
CHILDREN HOMS	SH05061	SH05062	SH05063	SH05064	SH05065														
EDUCATIONAL EST	SH05066	SH05067	-	-	-														
PCES OF DETENTC	SH05069	SH05070	SH05071	SH05072	SH05073														
DEFENCE ESTABS	SH05074	SH05075	-	-	-														
CIVILIAN VESSELS	SH05077	SH05078	-	-	-														
MISCEL. COMMUN	SH05080	SH05081	-	-	-														
MISCELLANEOUS	SH05083	SH05084	-	-	-														
Table SH07										MALE		FMLE							
Table SH08										OLD PERSONS ALONE	SH07001	SH07002							
Table SH09										OLD PERSONS IN HDS OF 2PERSONS-									
Table SH10										ONE OLD	SH08001	SH08002							
Table SH11										TWO OLD	SH08003	SH08004							
Table SH12										MARRIED COUPLE	SH08005	SH08006							
Table SH13										OTHERS-	SH08007	SH08008							
Table SH14										MALE OLDE	SH08009	SH08010							
Table SH15										MALE OLDE	SH08011	SH08012							
Table SH16										RELATIVES	SH08013	SH08014							
Table SH17										RESIDENT G	SH08015	SH08016							
Table SH18										VISITOR GU	SH08017	SH08018							
Table SH19										ADULTS	SH08019	SH08020							
Table SH20										ADULTS	SH08021	SH08022							
Table SH21										ADULTS	SH08023	SH08024							
Table SH22										ADULTS	SH08025	SH08026							
Table SH23										ADULTS	SH08027	SH08028							
Table SH24										ADULTS	SH08029	SH08030							
Table SH25										ADULTS	SH08031	SH08032							
Table SH26										ADULTS	SH08033	SH08034							
Table SH27										ADULTS	SH08035	SH08036							
Table SH28										ADULTS	SH08037	SH08038							
Table SH29										ADULTS	SH08039	SH08040							
Table SH30										ADULTS	SH08041	SH08042							
Table SH31										ADULTS	SH08043	SH08044							
Table SH32										ADULTS	SH08045	SH08046							
Table SH33										ADULTS	SH08047	SH08048							
Table SH34										ADULTS	SH08049	SH08050							
Table SH35										ADULTS	SH08051	SH08052							
Table SH36										ADULTS	SH08053	SH08054							
Table SH37										ADULTS	SH08055	SH08056							
Table SH38										ADULTS	SH08057	SH08058							
Table SH39										ADULTS	SH08059	SH08060							
Table SH40										ADULTS	SH08061	SH08062							
Table SH41										ADULTS	SH08063	SH08064							
Table SH42										ADULTS	SH08065	SH08066							
Table SH43										ADULTS	SH08067	SH08068							
Table SH44										ADULTS	SH08069	SH08070							
Table SH45										ADULTS	SH08071	SH08072							
Table SH46										ADULTS	SH08073	SH08074							
Table SH47										ADULTS	SH08075	SH08076							
Table SH48										ADULTS	SH08077	SH08078							
Table SH49										ADULTS	SH08079	SH08080							
Table SH50																			

Templates

Aletheia - [SAS_A_PAT1.xml] (Colour Image: 00773945.tif, B/W Image: 00773945_b.tif)

Home View Image Bounds Regions (F6) Text Lines (F7) Words (F8) Glyphs (F9) Dewarping Experimental

Save all changes Save Metadata Statistics Attributes Rotate left Rotate right Crop page Transform Select all Edit Zoom in Zoom out Full Page Colour B/W Analyse Page Auto Validation Layout Evaluation Quality Settings Interface Customisation About Updates User Guide Introduction Toolbar Open Example Aletheia Help

Pages Page Collection

SAS_A_PAT1.xml

New Page

<XML>

Add Page

Table TENURE BY TENURE

	HHS	PSNS	ROOMS
OWNERS OCCUPIED	89	283	518
RENTING IN BUSINESS	5	15	15
HOUSING BY EMPLOYMENT	6	15	29
RENTING FROM COUNCIL	5	16	21
RENTING FURNISHED	6	12	23
RENTING UNFURNISHED	148	662	746

Table DWELLINGS

DWELL TYPE	HHS	PSNS	ROOMS
1ST	231	228	231
2ND	12	18	55
3RD	22	22	53
4TH	22	22	75

Table CHILD ARRANGEMENTS

	ACD	HSDS	SIG	SHG
COLD WATER SHOWER	8	0	0	0
HOT WATER SHOWER	0	0	0	0
FIXED BATH SHOWER	6	0	0	0
WATER CIST SHOWER	0	0	0	0
ALL EXCLUSIVE	61	0	0	0

Table PERSONS BORN OUTSIDE UK

NATIONALITY	MALES	FEMALES
ENGLAND & WALES	7	6
INDIA & PAKISTAN	1	1
CEYLON	2	2
BRITISH & AFRICA	1	1
BR. & EXTRABRICA	1	1
BR. & CARIB. ISLANDS	1	1
MALTA	1	1
CYPRUS	1	1
OTHER COMMONWEALTH	1	1
FOREIGN AREAS	1	1

Table PERSONS ALONE

	MALES	FEMALES
ONE	120	120
MARRIED COUPLE	10	21
OTHERS - MALE IN ODER	2	0
FEMALE OLDER	0	2
BOTH MALE	2	2
BOTH FEMALE	0	0

Table NON-PRIVATE POPULATION - NOT HOTELS

NON-PERSONS	MALES	FEMALES	INMATES	NON-PERSONS
NHS ACUTE HOSPITALS	-	-	-	-
NHS CHRONIC HOSP.	-	-	-	-
NHS PSYCHIATRIC	-	-	-	-
NHS ISO ATTEN HOS	-	-	-	-
NHS OTHER HOSP.	-	-	-	-
OTHER MATERNITY	-	-	-	-
CHEM. PSYCHIATRIC	-	-	-	-
REFUGEE & EXCELEN	-	-	-	-
OTHER HOSPITALS	-	-	-	-
HOMES FOR AGED	-	-	-	-
HOMES FOR DISABLED	-	-	-	-
AGED AND DISABLED	-	-	-	-
CHILDREN'S HOMES	-	-	-	-
EDUCATIONAL ESTAB.	-	-	-	-
EX-PROSTITUTION	-	-	-	-
DEFENCE ESTAB.	-	-	-	-
CIVILIAN VSFCS	-	-	-	-
MISCELL. COMMUNAL	-	-	-	-
MISCELLANEOUS	-	-	-	-

Table SIZE OF UNDER 100 ROOMS

NO. OF HOTELS	1-10 ROOMS	11-20 ROOMS	21-30 ROOMS	31-40 ROOMS	41-50 ROOMS	51-60 ROOMS	61-70 ROOMS	71-80 ROOMS	81-90 ROOMS	91-100 ROOMS
PERSONS	1	1	1	1	1	1	1	1	1	1
ENTRANCE	1	1	1	1	1	1	1	1	1	1
ENTRANCE	1	1	1	1	1	1	1	1	1	1

Table NON-PRIVATE

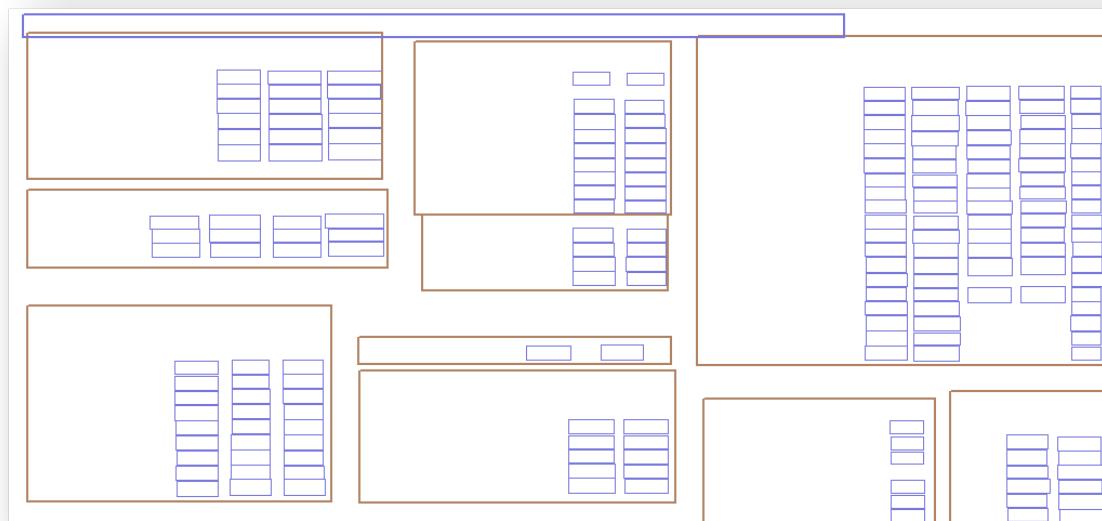
NON-PERSONS	MALES	FEMALES
1-10 RMS	1	1
11-14 RMS	1	1
15-24 RMS	1	1
25+	1	1

Run... 1268, 992

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

Templates

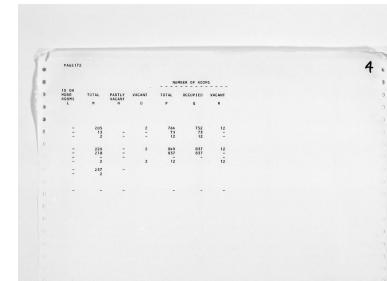
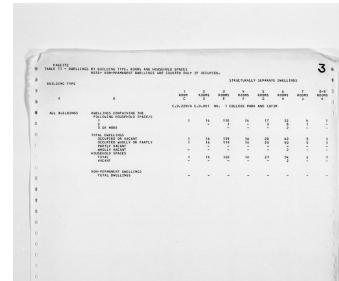
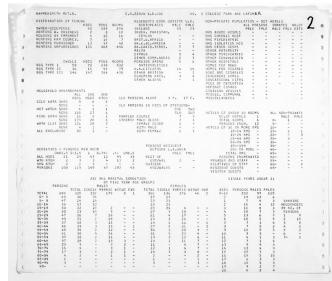
Je zwei Templates: Numerische Zellen und textuelle Zellen



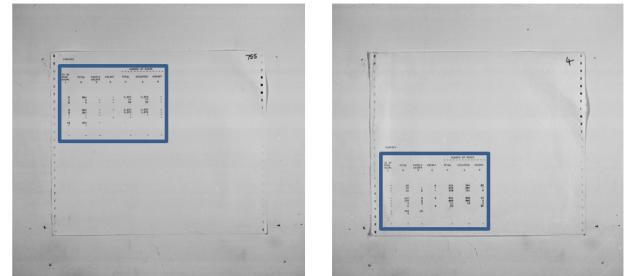
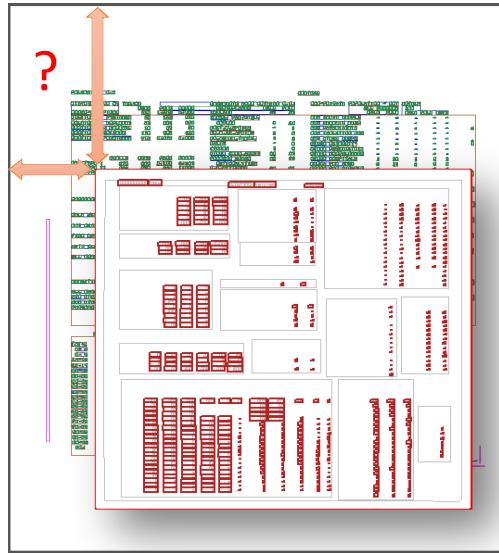
Wird wahrscheinlich
in Zukunft zu einem
Template vereinheitlicht.

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



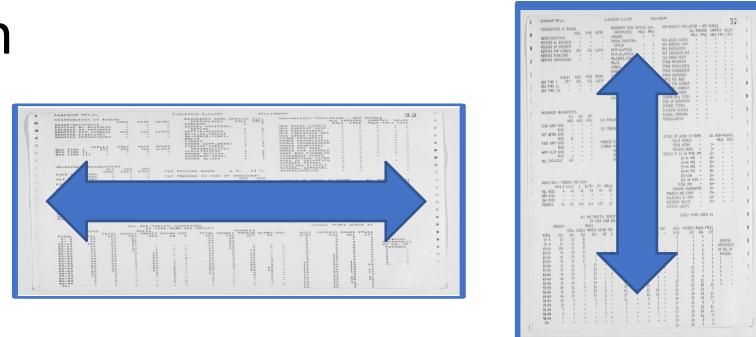
Template-Matching



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

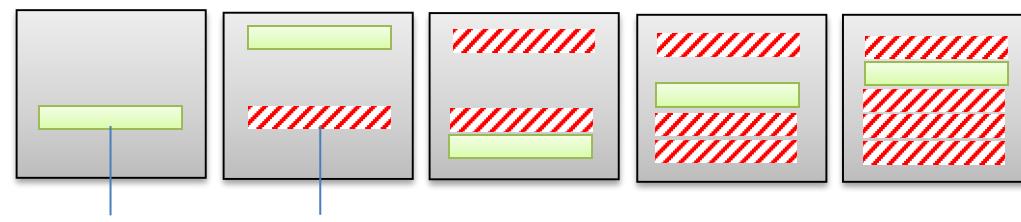
Skalierung

- 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



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 Regis- tration 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.	Staffs.	Uttlesford 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.	353/1	144	287
Abbots, Hanham	C.P.	Glouc.	Wormleaze R.D.	352/1	246	3,862
Abbots Haye	Loc.	Staffs.	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.	339/3	160	-

Area	Acreage	Population						Private households and dwellings. 1961					
		1951		1961				Private house-holds	Popula-tion in private house-holds	Struc-turally sepa-rate dwellings occupied	Rooms occupied	Density of occupation	
		Persons	Persons	Males	Females	Persons per acre	Percent of persons per room					Percent of persons per room	Percent of persons per room
b	c	d	e	f	g	h	i	j	k	l	m	n	
WORLDRIDGE	502,240	541,631	389,607	180,049	190,288	1,3	120,083	309,283	117,433	661,682	6,67	6,67	6,67

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)

SH02010

Data type xsd:integer ▾

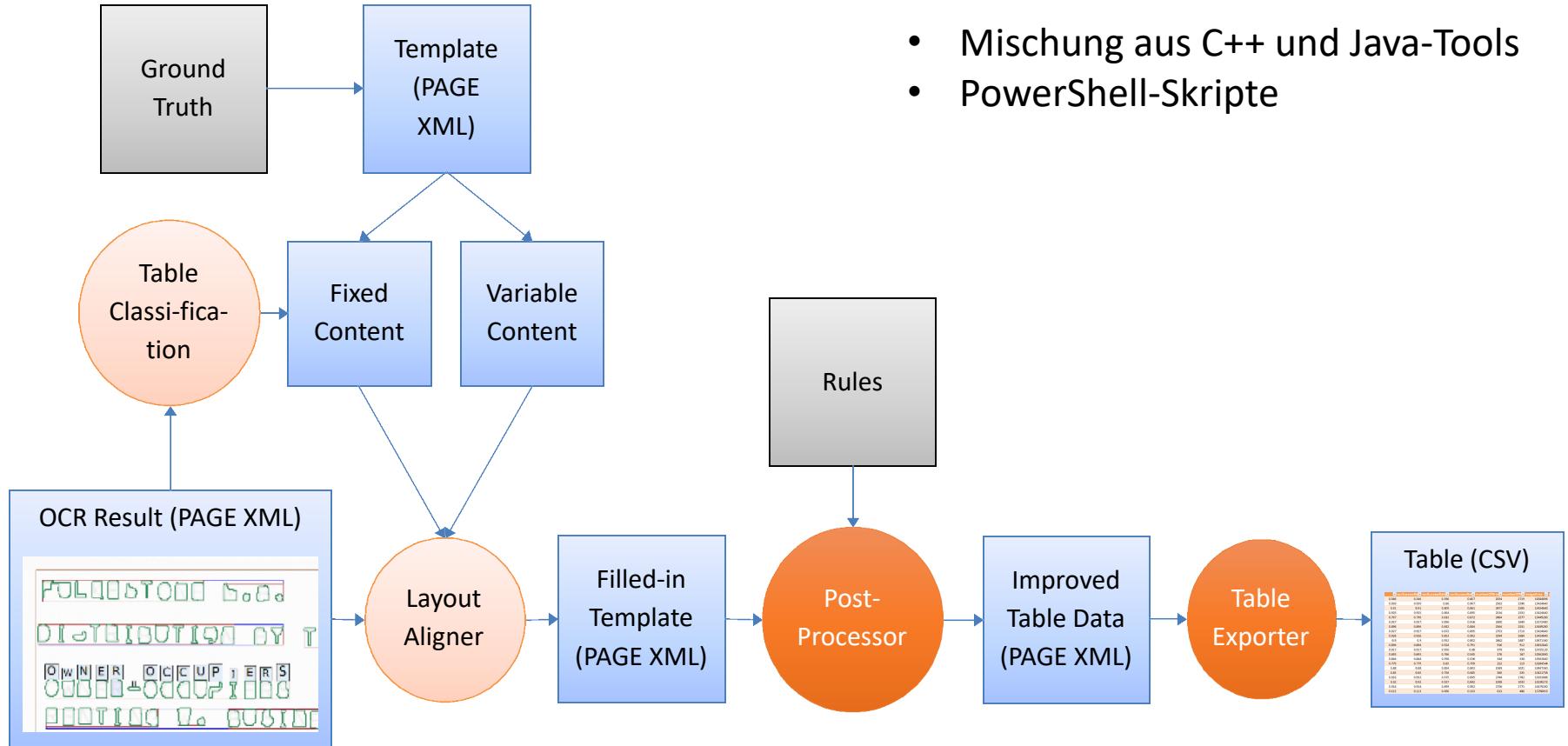
231

Zellen-ID

Datentyp

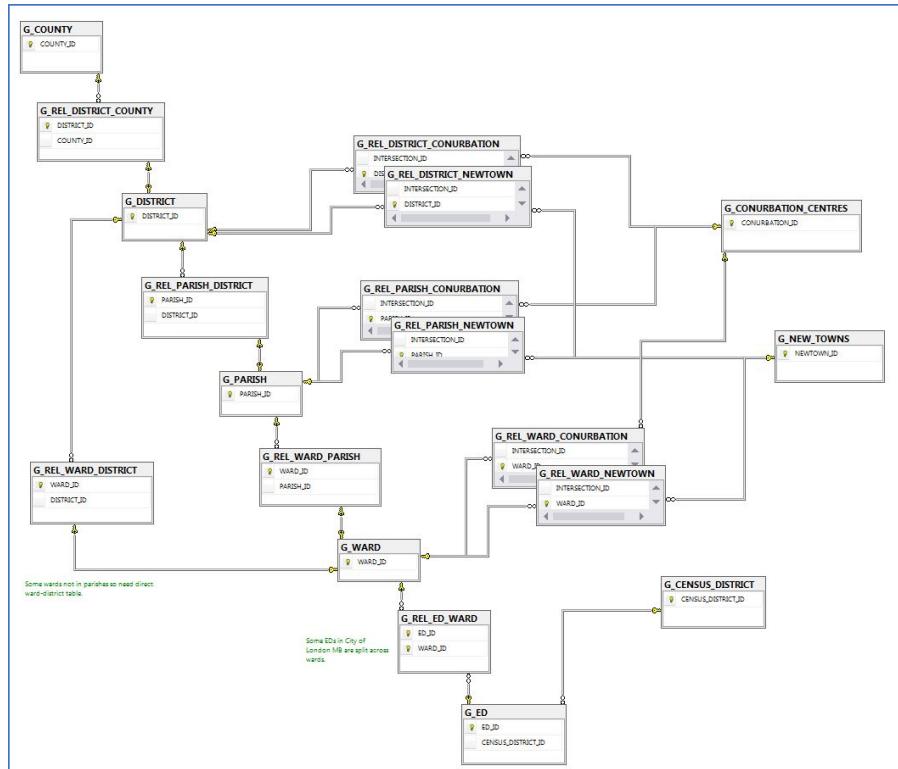
Bildausschnitt

Übersicht der “Pipeline”



Datenbank

- Import der extrahierten Daten
- Zensus-Modell



- Bisher Microsoft SQL Server
- Im Moment Umstieg auf MySQL

Qualitätskontrolle

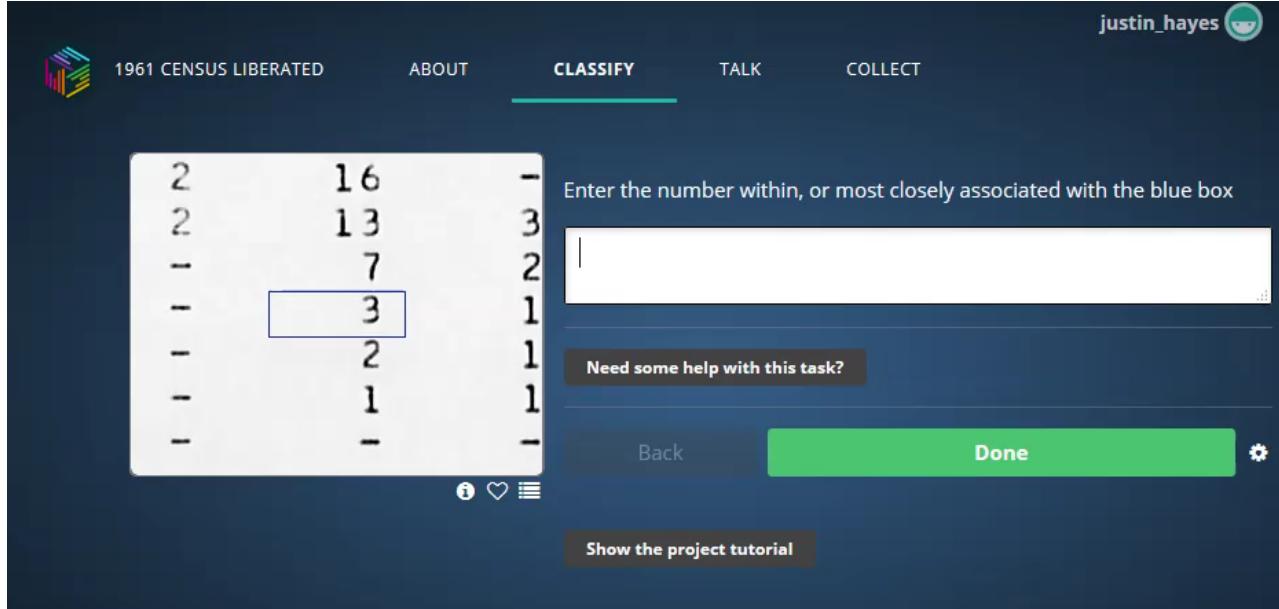
- Korrektheit der Zellen etwa 98%
- 2% falsche Zellen von Millionen Zellen ist viel!
- Confidence-Werte in der Pipeline (aber nur Anhaltspunkte)
- Besser: **Automatische Kontrolle**, um falsche Zellen zu finden
 - Nutzen von Datenredundanzen, um Fehler zu finden
 - Summen innerhalb einer Tabelle oder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

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



Nachkorrektur

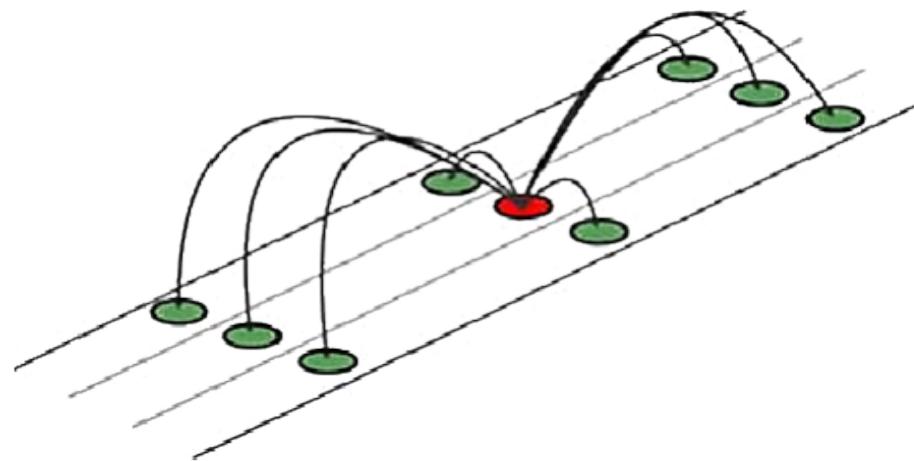
- Direkt oder mittels Crowd-Sourcing
- Zooniverse-Plattform
- Zelle für Zelle (mit Bildausschnitten vom originalen Scan)



The screenshot shows a Zooniverse classification task for the "1961 CENSUS LIBERATED" project. At the top, there are tabs for "CLASSIFY" (which is underlined in green), "ABOUT", "TALK", and "COLLECT". The user's name "justin_hayes" is displayed with a small profile icon. Below the tabs, a grid of numbers from a census form is shown. The number "3" in the fourth row is highlighted with a blue border. To the right of the grid, instructions say "Enter the number within, or most closely associated with the blue box" and a text input field contains the character "l". A help button says "Need some help with this task?". At the bottom, there are buttons for "Back", "Done", and a gear icon. A "Show the project tutorial" link is at the very bottom.

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