Data Warehousing and Data Mining (OLAP) for Accessing Archived Databases: A Case Study of the US 1880 Census

Richard Healey

Dept. of Geography

University of Portsmouth

The NAPP Dataset

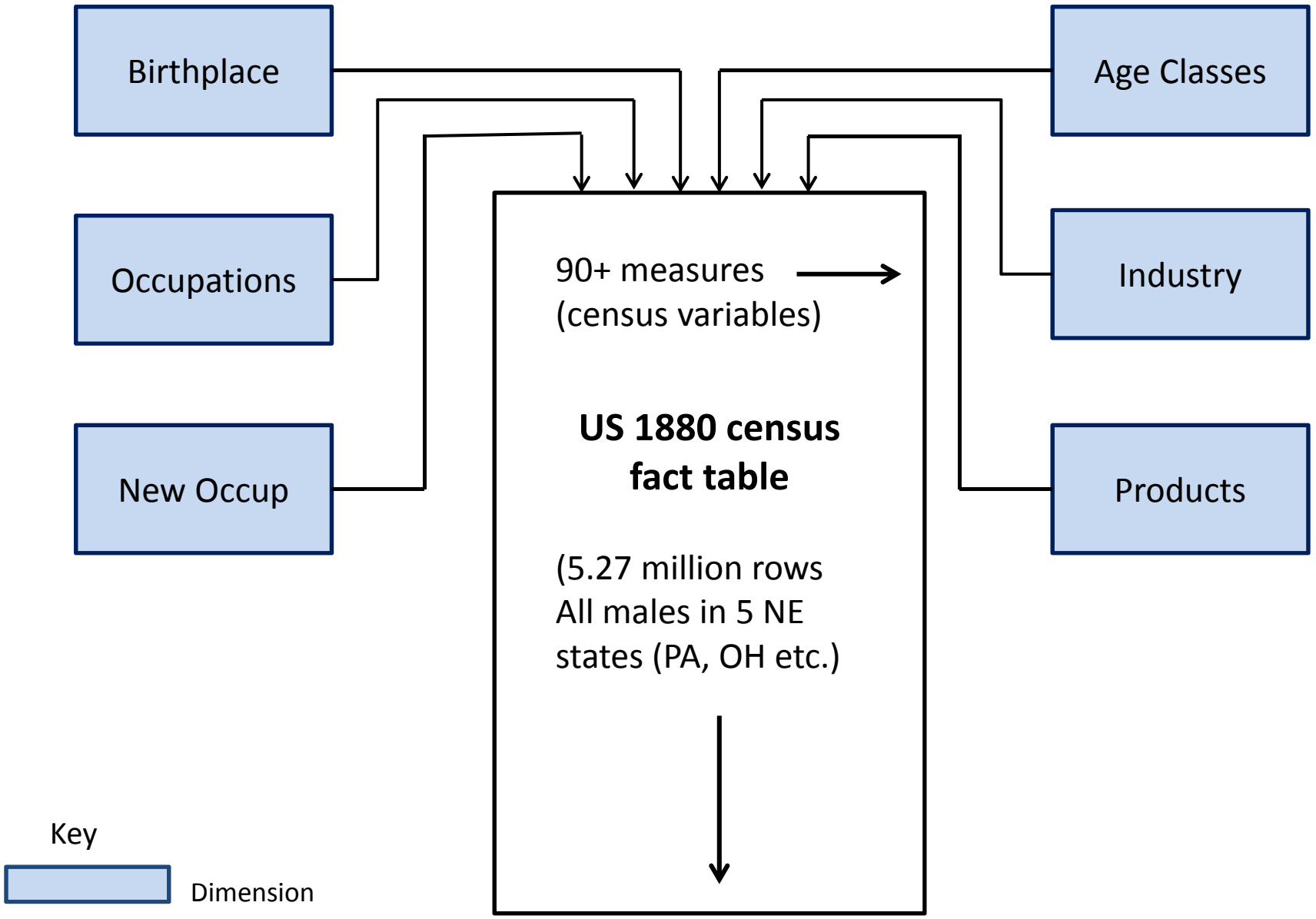
- Approx. 53 million individual person records are available from the US 1880 census
- Downloadable in bulk from the NAPP website
- Individual details of name, place of birth, age, occupation, parental birthplaces etc.
- Most fields converted to numeric codes
- First pilot - 164,000 heavy industrial workers chosen for the 67 counties of Pennsylvania
- Second ‘industrial strength’ data warehouse – 5.27 million records - entire male population of five states in the NE USA
- Recent transfer to supercomputer - enlargement under consideration

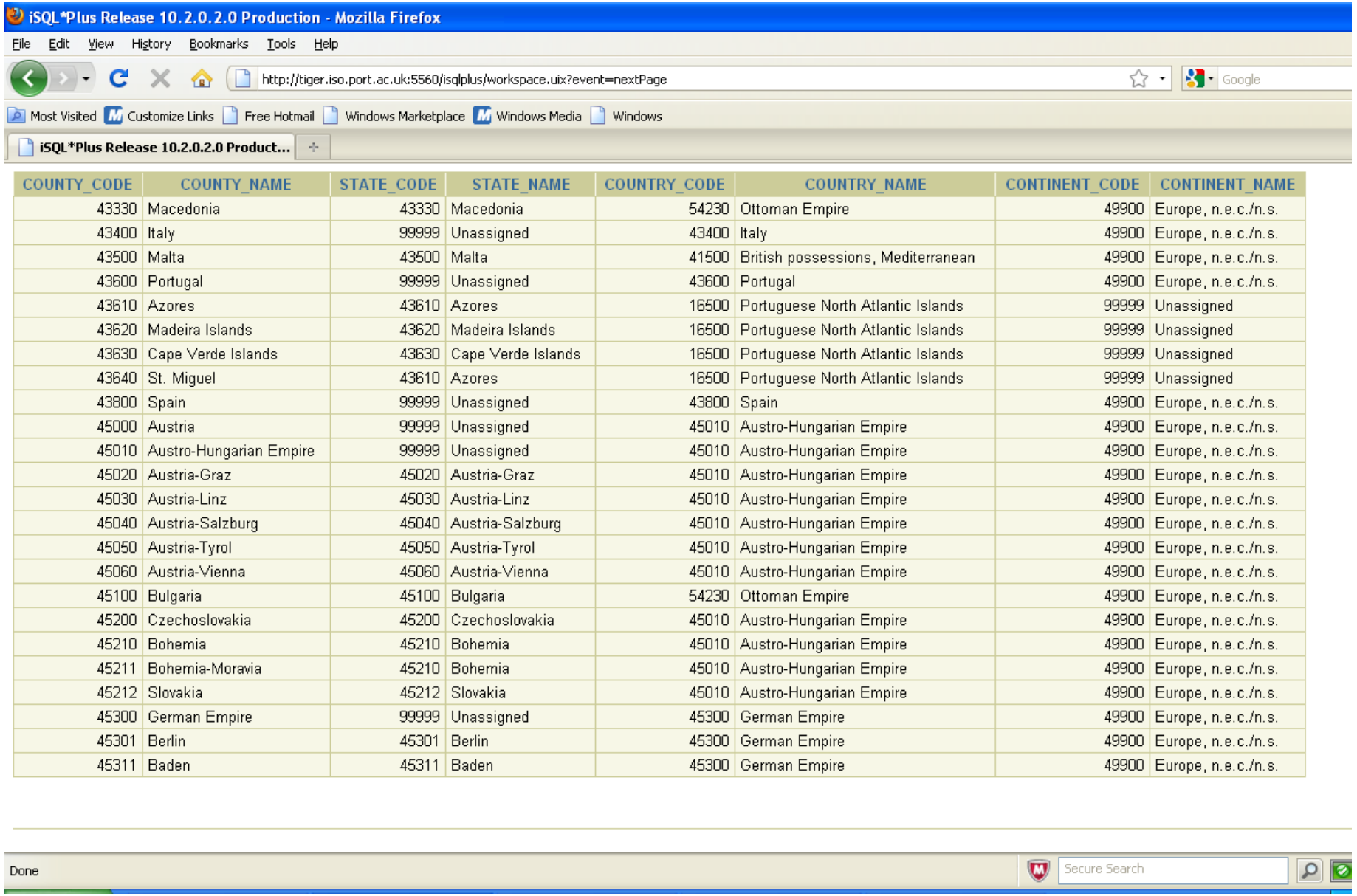
DW overview

- Designed for ‘decision support’ type applications not transaction processing
- Usually de-normalised for performance reasons
- Supports extremely large data volumes
- Employs specialised bitmap indexes to maximise performance
- Interfaces to Online-Analytical Processing (OLAP) tools to handle decision support queries

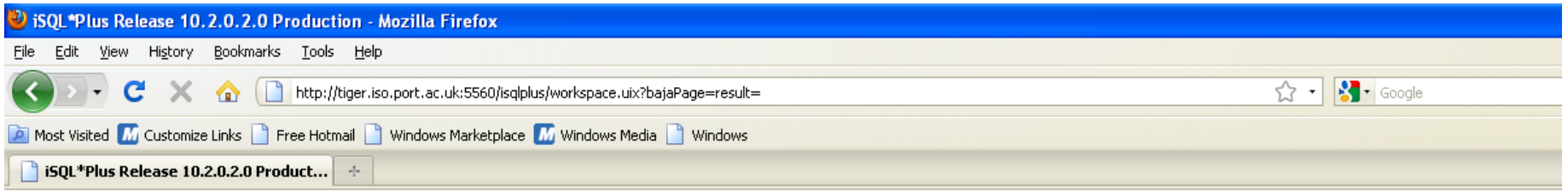
Specific Census DW structure

- *Fact/Cube* Table contains the coded individual person records
- The cube's *measures* are the census variables/attributes
- The cube is linked via database keys to a series of *dimensions*
- Dimensions are tabular representations of (often) hierarchical structures used to drive OLAP data (dis-)aggregation mechanisms

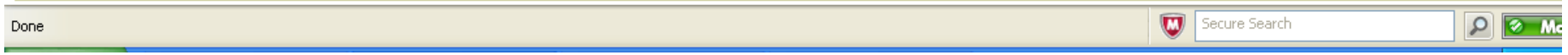

Birthplace
Occupations
New Occup
Age Classes
Industry
Products
90+ measures
(census variables)
US 1880 census
fact table
(5.27 million rows
All males in 5 NE
states (PA, OH etc.)
Key
Dimension

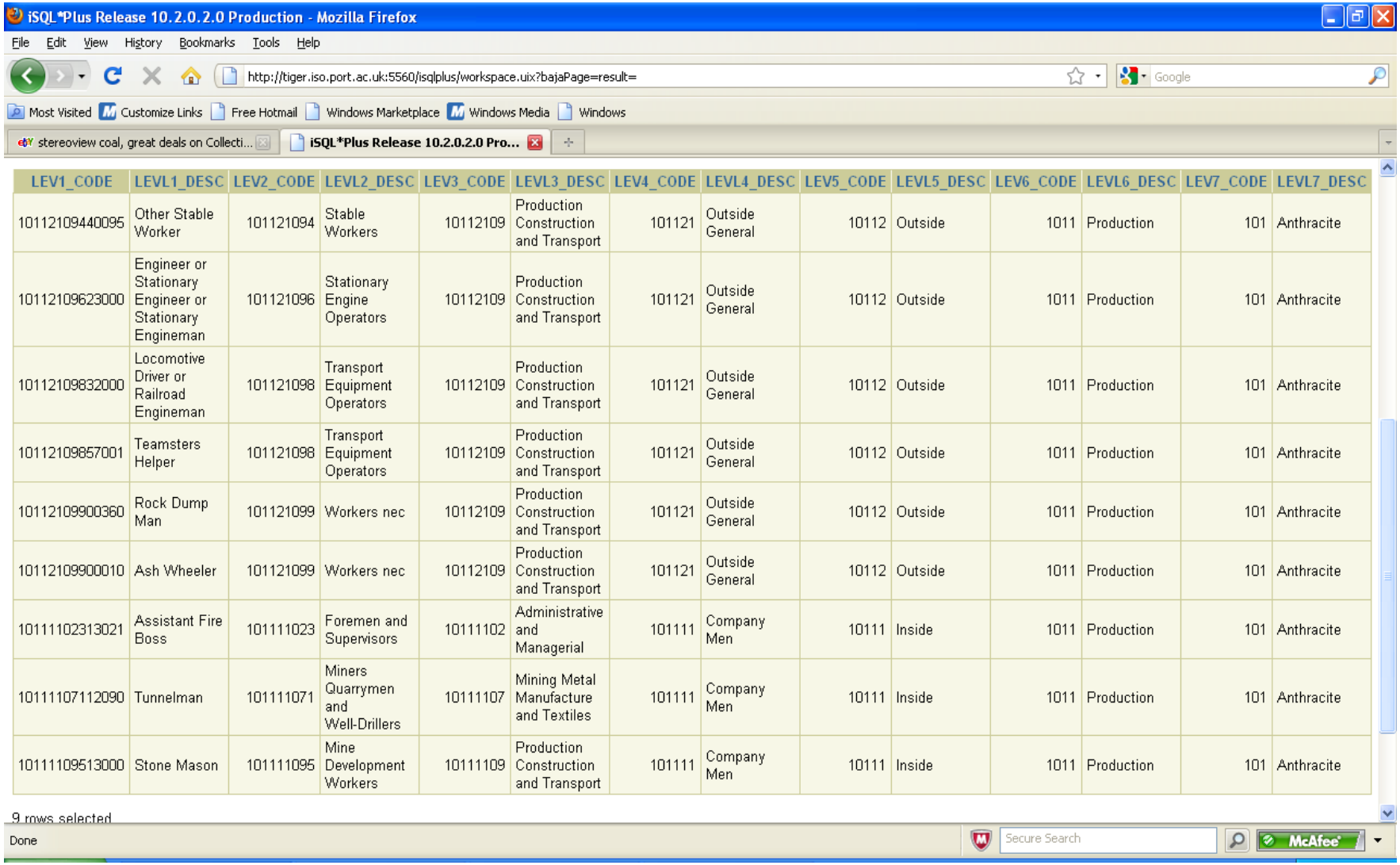


COUNTY_CODE	COUNTY_NAME	STATE_CODE	STATE_NAME	COUNTRY_CODE	COUNTRY_NAME	CONTINENT_CODE	CONTINENT_NAME
43330	Macedonia	43330	Macedonia	54230	Ottoman Empire	49900	Europe, n.e.c./n.s.
43400	Italy	99999	Unassigned	43400	Italy	49900	Europe, n.e.c./n.s.
43500	Malta	43500	Malta	41500	British possessions, Mediterranean	49900	Europe, n.e.c./n.s.
43600	Portugal	99999	Unassigned	43600	Portugal	49900	Europe, n.e.c./n.s.
43610	Azores	43610	Azores	16500	Portuguese North Atlantic Islands	99999	Unassigned
43620	Madeira Islands	43620	Madeira Islands	16500	Portuguese North Atlantic Islands	99999	Unassigned
43630	Cape Verde Islands	43630	Cape Verde Islands	16500	Portuguese North Atlantic Islands	99999	Unassigned
43640	St. Miguel	43610	Azores	16500	Portuguese North Atlantic Islands	99999	Unassigned
43800	Spain	99999	Unassigned	43800	Spain	49900	Europe, n.e.c./n.s.
45000	Austria	99999	Unassigned	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45010	Austro-Hungarian Empire	99999	Unassigned	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45020	Austria-Graz	45020	Austria-Graz	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45030	Austria-Linz	45030	Austria-Linz	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45040	Austria-Salzburg	45040	Austria-Salzburg	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45050	Austria-Tyrol	45050	Austria-Tyrol	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45060	Austria-Vienna	45060	Austria-Vienna	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45100	Bulgaria	45100	Bulgaria	54230	Ottoman Empire	49900	Europe, n.e.c./n.s.
45200	Czechoslovakia	45200	Czechoslovakia	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45210	Bohemia	45210	Bohemia	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45211	Bohemia-Moravia	45210	Bohemia	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45212	Slovakia	45212	Slovakia	45010	Austro-Hungarian Empire	49900	Europe, n.e.c./n.s.
45300	German Empire	99999	Unassigned	45300	German Empire	49900	Europe, n.e.c./n.s.
45301	Berlin	45301	Berlin	45300	German Empire	49900	Europe, n.e.c./n.s.
45311	Baden	45311	Baden	45300	German Empire	49900	Europe, n.e.c./n.s.



SUBSECTOR_VAL	SUBSECTOR_NAME	SECTOR_VAL	SECTOR_NAME	MAINDIV_VAL	MAINDIV_NAME
206	Metal mining	20	Mining	2	Mining
216	Coal mining	20	Mining	2	Mining
226	Crude petroleum and natural gas extraction	20	Mining	2	Mining
236	Nonmetallic mining and quarrying but not fuel	20	Mining	2	Mining
239	Mining not specified (1880)	20	Mining	2	Mining
306	Logging	30	Wood	3	Durable Goods Manufacturing
307	Saw- and planing- mills and mill work	30	Wood	3	Durable Goods Manufacturing
308	Miscellaneous wood products	30	Wood	3	Durable Goods Manufacturing
309	Furniture and fixtures	30	Wood	3	Durable Goods Manufacturing
316	Glass and glass products	31	Miscellaneous	3	Durable Goods Manufacturing
317	Cement concrete gypsum plaster	31	Miscellaneous	3	Durable Goods Manufacturing
318	Structural clay products	31	Miscellaneous	3	Durable Goods Manufacturing
319	Pottery and related products	31	Miscellaneous	3	Durable Goods Manufacturing
326	Misc nonmetallic mineral and stone	31	Miscellaneous	3	Durable Goods Manufacturing
336	Furnaces and steel works and rolling mills	32	Iron and Steel	3	Durable Goods Manufacturing
337	Other primary iron or steel industry	32	Iron and Steel	3	Durable Goods Manufacturing
346	Fabricated steel products	32	Iron and Steel	3	Durable Goods Manufacturing
338	Primary nonferrous industries	33	Nonferrous Metals	3	Durable Goods Manufacturing
347	Fabricated nonferrous metal	33	Nonferrous Metals	3	Durable Goods Manufacturing
348	Not specified metal industries	34	Unspecified Metals	3	Durable Goods Manufacturing
356	Agricultural machinery and tractors	35	Machinery	3	Durable Goods Manufacturing
357	Office and store machines and devices	35	Machinery	3	Durable Goods Manufacturing
358	Miscellaneous machinery	35	Machinery	3	Durable Goods Manufacturing
367	Electrical machinery or equipment and supplies	35	Machinery	3	Durable Goods Manufacturing





LEV1_CODE	LEVL1_DESC	LEV2_CODE	LEVL2_DESC	LEV3_CODE	LEVL3_DESC	LEV4_CODE	LEVL4_DESC	LEV5_CODE	LEVL5_DESC	LEV6_CODE	LEVL6_DESC	LEV7_CODE	LEVL7_DESC
10112109440095	Other Stable Worker	101121094	Stable Workers	10112109	Production Construction and Transport	101121	Outside General	10112	Outside	1011	Production	101	Anthracite
10112109623000	Engineer or Stationary Engineer or Stationary Engineman	101121096	Stationary Engine Operators	10112109	Production Construction and Transport	101121	Outside General	10112	Outside	1011	Production	101	Anthracite
10112109832000	Locomotive Driver or Railroad Engineman	101121098	Transport Equipment Operators	10112109	Production Construction and Transport	101121	Outside General	10112	Outside	1011	Production	101	Anthracite
10112109857001	Teamsters Helper	101121098	Transport Equipment Operators	10112109	Production Construction and Transport	101121	Outside General	10112	Outside	1011	Production	101	Anthracite
10112109900360	Rock Dump Man	101121099	Workers nec	10112109	Production Construction and Transport	101121	Outside General	10112	Outside	1011	Production	101	Anthracite
10112109900010	Ash Wheeler	101121099	Workers nec	10112109	Production Construction and Transport	101121	Outside General	10112	Outside	1011	Production	101	Anthracite
10111102313021	Assistant Fire Boss	101111023	Foremen and Supervisors	10111102	Administrative and Managerial	101111	Company Men	10111	Inside	1011	Production	101	Anthracite
10111107112090	Tunnelman	101111071	Miners Quarrymen and Well-Drillers	10111107	Mining Metal Manufacture and Textiles	101111	Company Men	10111	Inside	1011	Production	101	Anthracite
10111109513000	Stone Mason	101111095	Mine Development Workers	10111109	Production Construction and Transport	101111	Company Men	10111	Inside	1011	Production	101	Anthracite

Example Codes for Anthracite Mining Occupations

Custom Tabulation Engine - System infrastructure

- Uses ORACLE DW technology, where the DW sits inside a relational database, but deploys *extensive* metadata structures to manage the DW functions.
- The OLAP operations are implemented as SQL extensions that 'understand' the metadata and the cube/dimension table structures
- This is known as a relational OLAP implementation
- It creates a virtual multi-dimensional data cube that can be sliced, diced, aggregated (rolled up) and drilled down, using appropriately designed queries
- Suited to exploiting availability of individual level census data where there is a desire for wide-ranging custom aggregation/data tabulation
- System can be linked to other ORACLE modules – SPATIAL, TEXT, DATA MINING to form *analytical cascade*

Example SQL query with OLAP extension

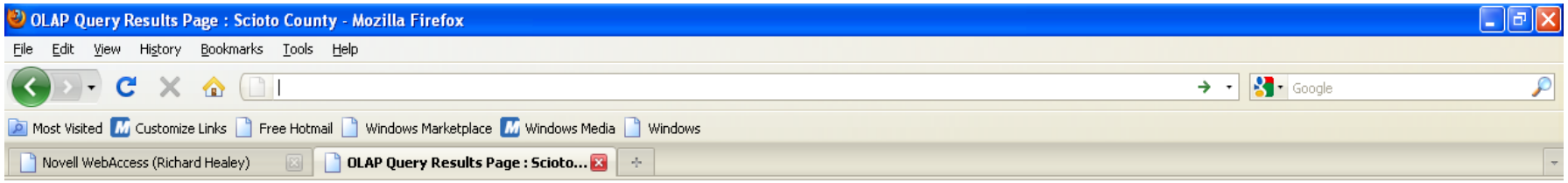
- select c.cv_county_code3,nvl(to_char(g.ctv_country_code),'Total'),
- nvl(a.nl_narrow_val_name,'Total'), count(c.low_age_val)
- from owb10_tar1.uscensus1880 c, owb10_tar1.geog_dim g, owb10_tar1.agegroups a
- where c.cv_county_code= g.cv_county_code
- and c.low_age_val=a.low_age_val
- and c.cv_county_code3 in (4200790, 4200690,4200030)
- group by c.cv_county_code3,
- cube(g.ctv_country_code,a.nl_narrow_val_name)

Pilot web-based system implementation

- Implemented in ORACLE PL/SQL with the following components :
  - OLAP query generation
  - OLAP query post-processing
  - Simple web interface (accesses dimensions/displays tables)
  - GIS analytical processing using ORACLE spatial functions
  - SVG graphics generator for map display and map-based OLAP query generation

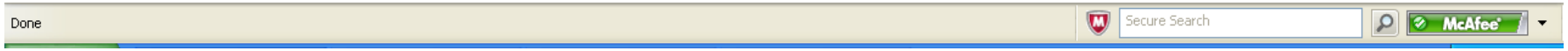
Insights Relevant to E-ARK

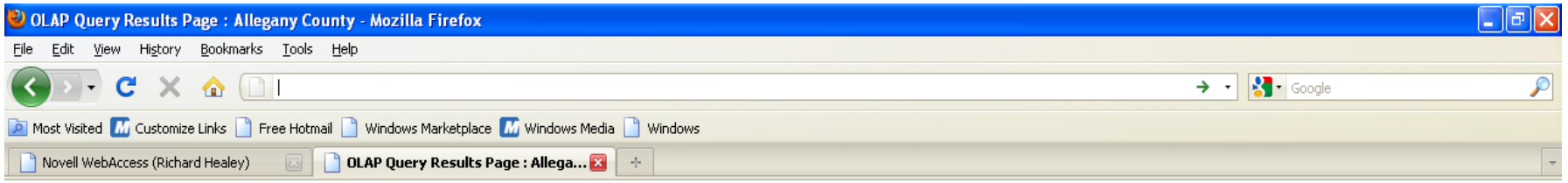
- Tight coupling of dimension and fact table keys removes problem of data mismatches (GSI/GSO principle)
- Dimensions are ‘mini-repositories’ of valuable structures for data standardisation across database snapshots and data tables from different sources (can be used outside DW also – e.g. occupations in B&O payrolls 1842-1857)
- Time dimension less interesting for census than, for example, business records – monthly payrolls etc., but such a general purpose dimension would apply across wide range of archived tables (as would geography, industry, occupation dimensions)
- Large ‘upfront’ investment in implementing dimensions but considerable payoff as archive grows



OLAP Query Results Page : Scioto County

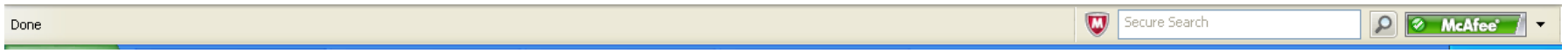
Birthplace	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100-104	105-109	110-114	115-119	120-124	125-129	Unknown	Total
At sea	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Austro-Hungarian Empire	0	0	0	0	0	1	0	1	0	2	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	6
Belgium	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3
Brazil	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
British West Indies, n.s.	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Canada	0	0	1	1	0	2	2	3	4	1	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	17
Denmark	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
France	0	0	0	2	2	8	9	6	12	18	12	9	12	8	11	2	3	0	0	0	0	0	0	0	0	0	1	115
German Empire	1	1	6	13	24	55	75	114	123	123	140	128	99	59	32	14	8	1	0	3	0	0	0	0	0	0	0	1019
Italy	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Mexico	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Missing/blank	2	1	1	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	10
Netherlands	0	0	0	0	0	1	0	2	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5
Russian Empire	0	0	0	1	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
Switzerland	0	1	0	1	0	2	4	7	2	3	1	7	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	32
United Kingdom, n.s.	5	7	9	5	7	18	32	39	38	51	45	39	32	29	17	6	4	0	0	0	0	0	0	0	0	0	0	383
United States, n.s.	2458	2315	2125	1747	1643	1214	849	720	550	440	316	279	209	172	103	74	35	14	1	0	1	0	0	0	0	0	1	15266

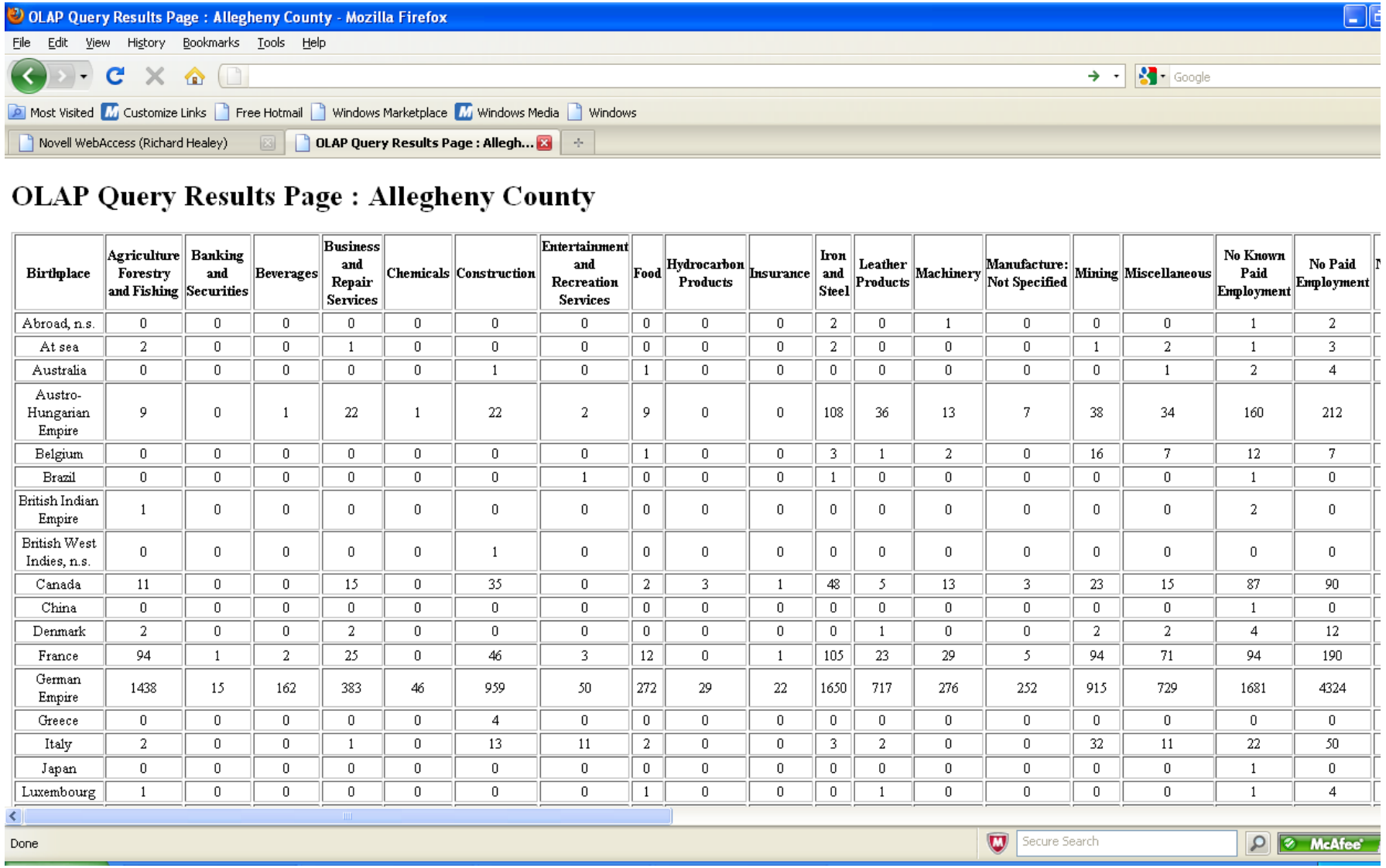




OLAP Query Results Page : Allegany County

Birthplace	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100-104	105-109	110-114	115-119	120-124	125-129	Unknown	Total
At sea	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Austro-Hungarian Empire	6	0	1	2	5	5	5	6	3	2	0	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	39
Brazil	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Canada	0	3	9	14	11	15	14	8	17	2	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96
Egypt	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
France	0	0	0	0	0	1	0	1	3	0	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	10
German Empire	5	7	19	18	33	62	93	69	94	133	112	81	74	51	41	15	8	3	0	0	0	0	0	0	0	0	0	918
Italy	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Missing/blank	2	0	2	1	4	4	0	0	3	1	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	23
Netherlands	0	0	0	0	2	1	4	3	1	1	2	0	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	19
Norway	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Russian Empire	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Spain	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Sweden	0	0	0	0	0	0	0	1	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Switzerland	0	0	0	1	1	1	0	0	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	7
Unassigned	0	0	1	1	0	1	0	1	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
United Kingdom, n.s.	9	47	170	202	216	243	309	337	276	194	213	115	140	57	40	19	9	2	0	0	0	0	0	0	0	0	0	2598
United States, n.s.	3028	2748	2180	1635	1537	1118	821	671	446	369	282	200	187	114	74	47	23	1	1	0	0	0	0	0	0	0	1	15483
Unknown	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1



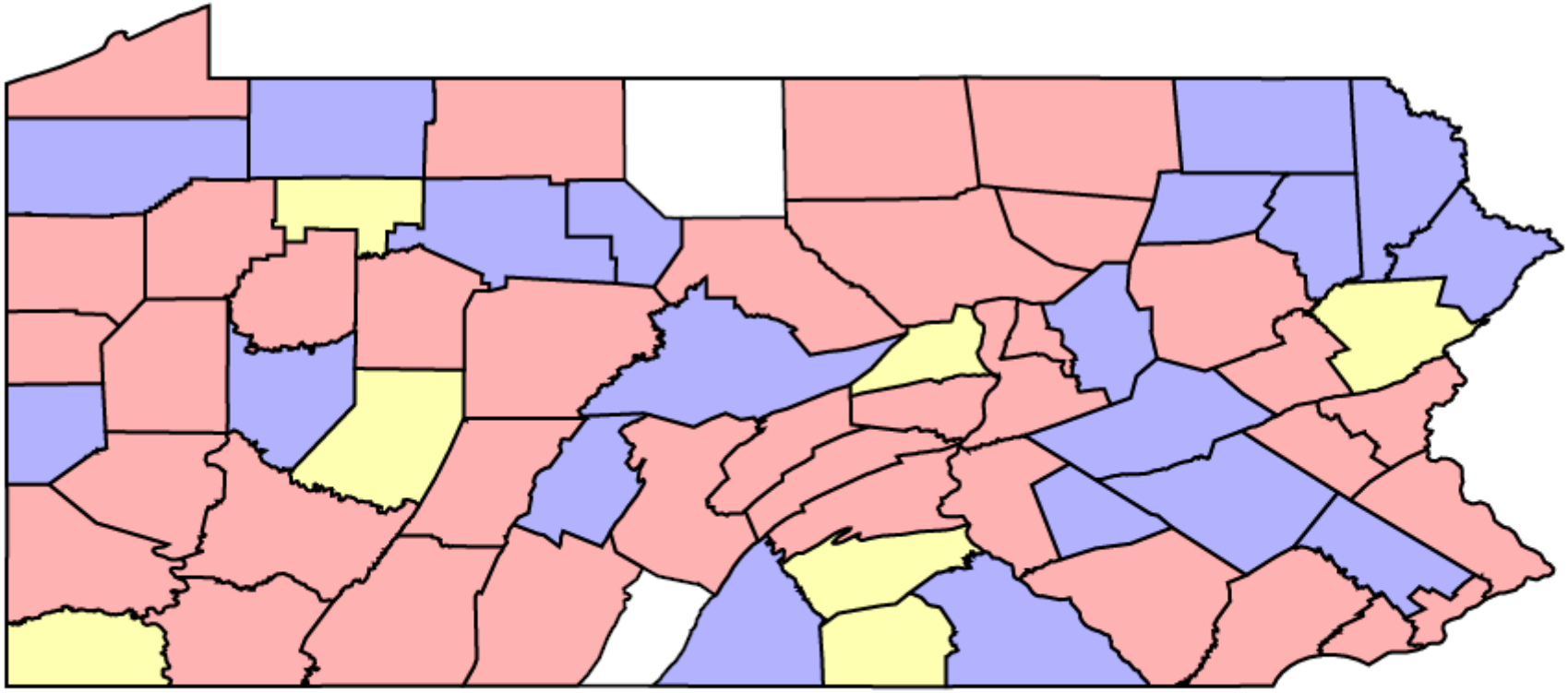


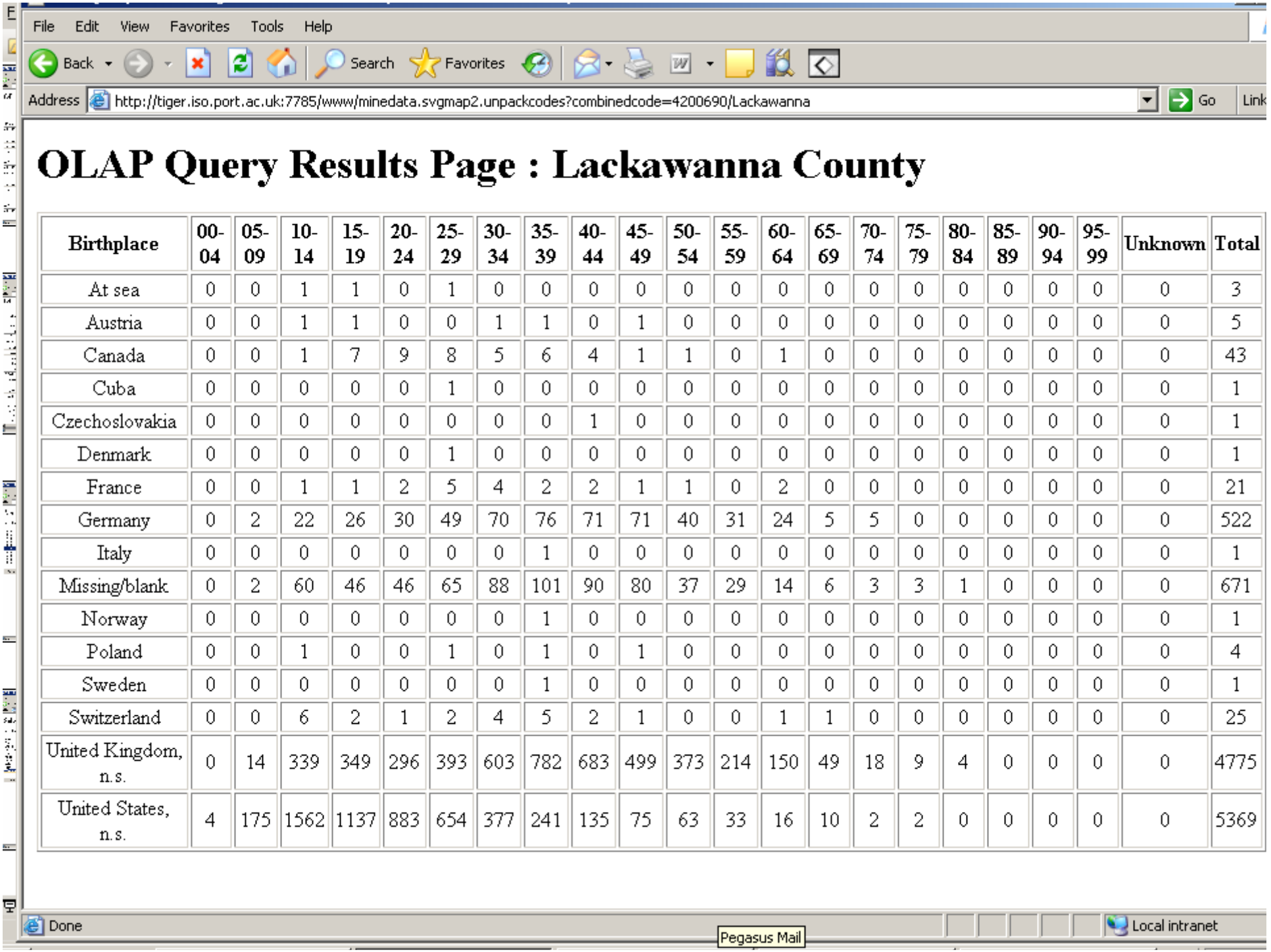
OLAP Query Results Page : Allegheny County

Birthplace	Agriculture Forestry and Fishing	Banking and Securities	Beverages	Business and Repair Services	Chemicals	Construction	Entertainment and Recreation Services	Food	Hydrocarbon Products	Insurance	Iron and Steel	Leather Products	Machinery	Manufacture: Not Specified	Mining	Miscellaneous	No Known Paid Employment	No Paid Employment
Abroad, n.s.	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	1	2
At sea	2	0	0	1	0	0	0	0	0	0	2	0	0	0	1	2	1	3
Australia	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	2	4
Austro-Hungarian Empire	9	0	1	22	1	22	2	9	0	0	108	36	13	7	38	34	160	212
Belgium	0	0	0	0	0	0	0	1	0	0	3	1	2	0	16	7	12	7
Brazil	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0
British Indian Empire	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
British West Indies, n.s.	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Canada	11	0	0	15	0	35	0	2	3	1	48	5	13	3	23	15	87	90
China	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Denmark	2	0	0	2	0	0	0	0	0	0	0	1	0	0	2	2	4	12
France	94	1	2	25	0	46	3	12	0	1	105	23	29	5	94	71	94	190
German Empire	1438	15	162	383	46	959	50	272	29	22	1650	717	276	252	915	729	1681	4324
Greece	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
Italy	2	0	0	1	0	13	11	2	0	0	3	2	0	0	32	11	22	50
Japan	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Luxembourg	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	4

Pennsylvania 1880 Census

Percentage of UK-born Heavy Industrial Workers in 25-29 Age Bracket





OLAP Query Results Page : Lackawanna County

Birthplace	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	Unknown	Total
At sea	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
Austria	0	0	1	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	5
Canada	0	0	1	7	9	8	5	6	4	1	1	0	1	0	0	0	0	0	0	0	0	43
Cuba	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Czechoslovakia	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
Denmark	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
France	0	0	1	1	2	5	4	2	2	1	1	0	2	0	0	0	0	0	0	0	0	21
Germany	0	2	22	26	30	49	70	76	71	71	40	31	24	5	5	0	0	0	0	0	0	522
Italy	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Missing/blank	0	2	60	46	46	65	88	101	90	80	37	29	14	6	3	3	1	0	0	0	0	671
Norway	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Poland	0	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	4
Sweden	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Switzerland	0	0	6	2	1	2	4	5	2	1	0	0	1	1	0	0	0	0	0	0	0	25
United Kingdom, n.s.	0	14	339	349	296	393	603	782	683	499	373	214	150	49	18	9	4	0	0	0	0	4775
United States, n.s.	4	175	1562	1137	883	654	377	241	135	75	63	33	16	10	2	2	0	0	0	0	0	5369