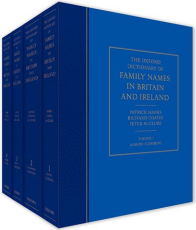
Pre-processing Large Resources for Family Names Research

Adam Rambousek

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
xrambous@fi.muni.cz
http://deb.fi.muni.cz
deb@fi.muni.cz

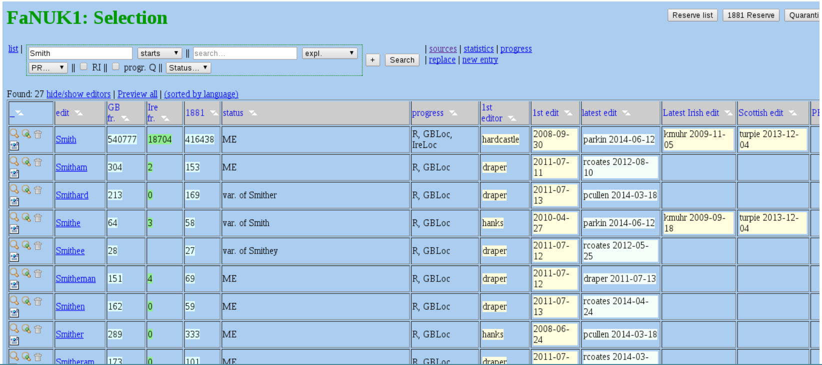
Introduction

- Family Names of the United Kingdom
- → Family Names in Britain and Ireland
- detailed investigation of the origins, history, geographical distribution of the surnames in the United Kingdom
- University of the West of England
- Arts and Humanities Research Council grants
  - 2010-2014: research of 45,000 most frequent surnames
  - 2014-2016: update and enlarge to 60,000 surnames
- phase 1 results published by OUP, November 2016
  - The Oxford Dictionary of Family Names in Britain and Ireland



Software tools

- editing and management tools based on the DEB platform
- authors, consultants, progress tracking, statistics, bibliography
- names collections: main FaNBI, FaNUK1, reserve, rare established, quarantine



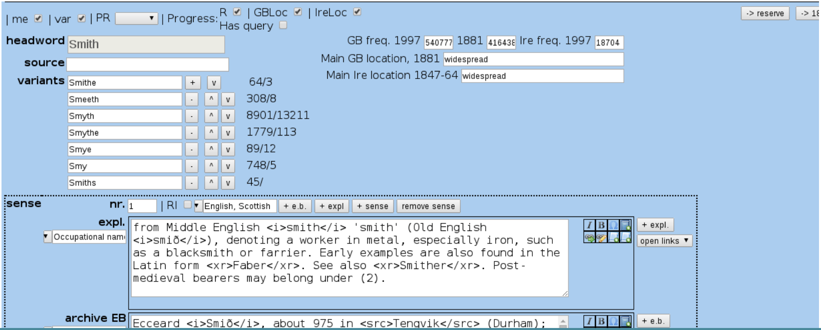
FaNUK1: Selection

Found: 27 hide/show editors | Preview all | (sorted by language)

_	edit	GB fr.	Ire fr.	1881	status	progress	1st editor	1st edit	latest edit	Latest Irish edit	Scottish edit	PI
	Smith	540777	18704	416438	ME	R, GBLoc, IreLoc	hardcastle	2008-09-30	parkin 2014-06-12	kmuhr 2009-11-05	turpie 2013-12-04	
	Smitham	304	2	153	ME	R, GBLoc	draper	2011-07-11	rcoates 2012-08-10			
	Smithard	213	0	169	var. of Smither	R, GBLoc	draper	2011-07-13	pcullen 2014-03-18			
	Smithe	64	3	58	var. of Smith	R, GBLoc	hanks	2010-04-27	parkin 2014-06-12	kmuhr 2009-09-18	turpie 2013-12-04	
	Smithee	28		27	var. of Smithey	R, GBLoc	draper	2011-07-12	rcoates 2012-05-25			
	Smitheman	151	4	69	ME	R, GBLoc	draper	2011-07-12	draper 2011-07-13			
	Smithen	162	0	59	ME	R, GBLoc	draper	2011-07-13	rcoates 2014-04-24			
	Smither	289	0	333	ME	R, GBLoc	hanks	2008-06-24	pcullen 2014-03-18			
	Smitheram	173	0	101	ME	R, GBLoc	draper	2011-07-	rcoates 2014-03-			

Software tools

- editing in "clusters", data validations (cross-references, bibliography,...)
- access resources for name (IGI, Poll Tax, Patent Rolls, Chancery Proceedings, The National Archives,...)
- requests, comments, further research...



| me ☑ | var ☑ | PR ▾ | Progress: R ☑ | GBLoc ☑ | IreLoc ☑
Has query ☐

headword Smith

source

GB freq. 1997 540777 1881 416438 Ire freq. 1997 18704
Main GB location, 1881 widespread
Main Ire location 1847-64 widespread

variants	Smithe	64/3
	Smeeth	308/8
	Smyth	8901/13211
	Smythe	1779/113
	Smye	89/12
	Smy	748/5
	Smiths	45/

sense nr. 1 | RI ☐ English, Scottish

expl. Occupational nam

```
from Middle English <i>smith</i> 'smith' (Old English
<i>smið</i>), denoting a worker in metal, especially iron, such
as a blacksmith or farrier. Early examples are also found in the
Latin form <xr>Faber</xr>. See also <xr>Smither</xr>. Post-
medieval bearers may belong under (2).
```

archive EB

```
Ecceard <i>Smið</i>, about 975 in <src>Tengvik</src> (Durham);
```

Frequency lists

- UK: 1881 census report, 1997 and 2011 statistical data
- Ireland: 1997 and 2008 statistical data
- dirty data:
  - spelling errors, invalid characters
  - all names in uppercase (MCGAFFIN vs McGaffin, DEBONO vs De Bono)
  - Irish and Scottish names not normalized (MACDOUGAL vs MCDOUGAL vs M'DOUGAL, OBRIAN vs O'BRIAN vs O'BRIAN vs O BRIAN)

Frequency lists – processing

- 0: detect Mc and O variants, lexicographers decide correct spelling (exceptions e.g. Mach, Mackarel)
- 1: use the list of variants to normalize names and sum frequencies
- 2: capitalize names (exceptions e.g. Van der Merwe, De Lisle)
- 3: normalize feminine forms (e.g. Sikorski/Sikorska), join to masculine

International Genealogical Index (IGI)

- compiled by the Mormons
- worldwide records extracted from the parish archives and similar sources, or submitted by the members of the Church
- for FaNBI – original database records for the Great Britain
  - 188,043,185 records (plain text files)
  - unreadable books, different spellings by each transcriber, spelling mistakes etc.

IGI record

(batch identification, event date, event place, event type, year, first name, surname, role, gender)

Ivy-church, Kent, England|26 Nov 1612|Ivy-Church, Kent, England|Marriage|1612|Alice|Darter|Bride|Female

International Genealogical Index – processing

- delete obvious mistakes (e.g. English cities in France)
- standardize county names
- extract place names for each county, distributed to volunteers to check spelling and correct city-county
- fix place names in IGI, delete records with incorrect place names
- detect (near) duplicate records
  - first name, surname, date, town, county, and event type
- 72,187,630 clean records

International Genealogical Index – enrich database

- add historical evidence automatically to FaNBI
- for each surname, extract IGI records for each century and most prominent county
- 40,274 surnames enriched
- all clean records available for researchers

IGI record converted to include in database

Bletsoe, Bedford, England|05 Sep 1629|Bletsoe, Bedford, England|Christening|1629|John|Darter|Principal's Father|Male

→

John <sn>Darter</sn>, 1629 in <src>IGI</src> (Bletsoe, Beds)

Fiants

- *The Irish Fiants of the Tudor sovereigns during the reigns of Henry VIII, Edward VI, Philip & Mary, and Elizabeth I*
- court warrants from Ireland, 1521-1603
- numbered records, official language
- Word documents, results of OCR

Fiants (Henry VIII)... record

213 (350). Pardon to Edward Nugent, of Stonehall or of Mylcastell, gentleman or horseman; especially for the murder of Edmund Nugent, of Multyfernane, gentleman.—30 June, xxxiii.
(Cal. P.R., p. 72, art. 72.)

Fiants – processing

- convert to XML
- extract each record to separate XML entry
- enhance metadata (e.g. convert date from regnal years)
- fix common OCR mistakes
- mark place names (list from IGI clean-up)
- mark occupation where possible
- mark first name, surname

Fiants (Edward VI)... record converted to include in database

846. Pardon to William Powell, of Dublin, soldier; especially for the death of John Randall.—2 October, v. (Cal. P.R., p. 244, art. 98.)
→
William <sn>Powell</sn>, soldier, 1551 in <src>Fiants Edward</src> $846 (Dublin);

Conclusions

- preprocessing and extraction of data in unified format
  - better overview of the name origin and distribution
  - rich information and historical evidence
  - faster lexicographic work
- tools and methodology adapted for *Dictionary of American Family Names (2nd edition)*