

## Thesaurus of British Surnames

### Minutes of Meeting of Steering Group 21/3/02 at SoG

#### Present:

Ian Galbraith, Peter Christian, Kevin Schürer, Steve Archer, John Dawson, Ben Laurie  
Apologies: David Thomas, Louise Craven

#### 1. TOBS Web Site

1.1 Site now up at <http://www.data-archive.ac.uk/surnames/> Members of the SG are asked to email KS with any suggestions.

ALL

1.2 Any documents for distribution should be emailed to KS, who will put them on the site. Emails to the mail list are limited as regards attachment size, so attachments should generally not be emailed.

1.3 KS will arrange for part of the site to be passworded, for access only by SG members, and he will distribute the password to SG members.

KS

1.4 All documents marked “draft” sent to KS will automatically be treated as non-public and placed in the passworded section of the site. (Unless Kevin is explicitly instructed otherwise, non-draft docs. will be assumed to be “public”.)

#### 2. Funding & Resources

2.1 KS has looked at the Marc Fitch and AHRB options, and talked to the Univ. of Sheffield.

2.1.1 AHRB (<http://www.ahrb.ac.uk/>): there doesn't appear to be a suitable scheme for the TOBS project.

2.1.2 Marc Fitch: Two possible categories: supported publications; research.

KS had spoken to a member of the committee. Awards are not generally more than £10-15K. They have a meeting in a few weeks and the next meeting is not for about six months, so we are too late to apply for now, but have time to consider an application for the next meeting.

2.1.3 Sheffield: Vaguely interested in the computational aspects, eg pattern matching, but no money. Might collaborate in future.

#### 2.2 Genealogy.com & Ancestry.com

BL has yet to talk to them.

BL

2.3 Univ. of Newcastle: Ben has spoken with Brian Randall, who is interested and has had research students working on this topic. He is now retired but maintains his interest.

2.4 MIMAS (<http://www.mimas.ac.uk/>): Ben will speak with Keith Cole, deputy director.

BL

2.5 Comment: No obvious source yet for funding for a project manager.

#### 3 PID

On hold till next meeting. Enough work to be going on with meantime.

#### 4. Source Data

4.1 Metadata for describing datasets

IG had distributed a “first pass” note for XML-based metadata. KS will put a copy of the spec used by the UK Data Archive on the TOBS web site.

KS

The TEI (Text Encoding Initiative) standard (<http://www.tei-c.org/>) may also have relevant components. IG will attempt to include relevant material from the above and from the TEI in a new draft of his note.

IG

It was agreed that the metadata for complete datasets could, where appropriate, be carried down to individual record level, and the record format should be so specified.

It was agreed to hold off obtaining additional datasets (from third parties) until the metadata required was finalised. It was anticipated that for such datasets the metadata could, in practice, be simply entered into a form: XML structures would ultimately be used for machine-machine transfer.

## 5. Methodology

### 5.1 What is a variant?

PC raised the question of what is a variant: spelling variants, certainly, but what about structural and etymological variants? Should these be included? You can identify morphological elements such as MC/MAC, ...SON, AP, O'. Should surnames which have added elements be treated as variants of the “root” surname? IG suggested that MC/MAC, etc, surnames with and without these elements not be considered variants. In practice researchers can be advised that such elements may have been added to or dropped from the root surname, but that the thesaurus should group root surnames separately from those with added patronymic and similar elements.

Comment: This issue needs a formal decision, to be included in the PID.

ALL?

### 5.2 Consonantal clusters

PC has been working on an analytical approach based upon consonantal clusters, as opposed to syllabic analysis; vowels are ignored. He feels that this is promising; there are relatively few consonantal clusters in English names. He is using the 1881 census, but would like frequency tables so he can isolate and work on the commonest names. KS will provide these tables, but commented that “his” surnames have spaces and apostrophes removed (eg DE ATH -> DEATH, O'BRIEN -> OBRIEN), and that there are no double-barrelled names.

KS

PC will continue to work on this approach and report on his findings.

PC

5.3 Variant surnames can be defined via algorithmic analysis and/or from documentary evidence and/or from consensus. Record structure has to accommodate this.

## 6. Database definition

What should go in the database?

### 6.1 Input data

Metadata at dataset level and optionally at record level.

### 6.2 Output data

BL suggests:

surname 1, surname 2, type of variant, authority, restriction on locality (optional; eg Yorkshire only), score

Thus larger or smaller (looser or tighter) groups of surnames can be generated on-the-fly depending on whether type of variant, restriction on locality and score value are selected.

(Comment. Since surname variants would be in pairs, the group SMITH, SMYTH, SMYTHER, for example, would be generated from three sets of pairs: SMITH, SMYTH; SMITH, SMYTHER; SMYTH, SMYTHER.)

7. Next meeting

Thursday, 9 May 2002, 1400, at Society of Genealogists, 9 Dallington Street, London.