

# BRITISH SURNAMES THESAURUS PROJECT

## MINUTES OF INAUGURAL MEETING

Held at the Public Record Office,

Tuesday 4 December 2001, 10.30 am, Conference Room C

### List of delegates

Hazel Anderson (National Archives of Scotland, SCAN Testaments Project Team Leader)  
Steve Archer (Federation of Family History Societies, National Burials Index)  
Richard Baker (Institute of Heraldry and Genealogical Studies)  
John Challis (Image Partners Ltd, Consultant)  
Peter Christian (Goldsmiths College)  
Else Churchill (Society of Genealogists, Genealogy Officer)  
Stella Colwell (PRO, Family Records Centre)  
Louise Craven (PRO, Authority Files Editor)  
David Crook (PRO, Research and Editorial Services Department)  
John Dawson (University of Cambridge, Computing Service)  
Ian Galbraith (Origins.net, Chairman) [chair]  
Graham Hart (FreeBMD, Co-founder and Techie)  
Chris Kutler (PRO, Systems Developer, e-Access)  
Ben Laurie (FreeBMD, Founder and Techie)/  
Camilla von Massenbach (FreeBMD, Co-founder)  
Geoff Mawlam (Genealogical Society of Utah)  
Peter Ruthven-Murray (Scottish Association of Family History Societies, Chairman)  
Professor Kevin Schurer (University of Essex, Department of History; Director, UK Data Archive)  
Peter Seaman (PRO, archivist, e-Access)  
David Squire (Society of Genealogists)  
Niall Taylor (National Archives of Scotland, SCAN Systems Analyst)  
David Thomas (PRO, Head of e-Access)

### ACTION

#### 1 Why a Surname Thesaurus?

- 1.1 Ian Galbraith (IG) welcomed all delegates and said how important he thought it was for the British Surnames Thesaurus Project to go ahead. David Thomas (DT) reciprocated and said this was an exciting project.

#### 2 The UK Data Archive (UKDA).

- 2.1 Kevin Schurer (KS), Director of the UK Data Archive (UKDA) at the University of Essex explained that The British Surnames Thesaurus Project could be hosted by the UK Data Archive. UKDA had been established for 35 years and was funded to provide a service for Higher Education, part of which included the History Data Service. The provision of access to individuals and bodies from the non-higher education sector required no

broadening of the UKDA remit: users were normally required to be registered, but some information was available to all.

2.2 KS had a personal interest in surnames, having been asked by the Science Museum, in the early 1990s to provide the research underlying the interactive display of British surnames, between 1881 and 1996. The display was ~~was~~ presented in the Wellcome Galleries

2.3 A lively discussion ensued concerning the number of unique surnames involved.

### 3 Surname problems in the National Burials Index (NBI) project.

3.1 Steve Archer (SA) introduced problems which had arisen in the National Burials Index (NBI) project. The project had originally intended to produce a searchable CD index of surnames. SA had used the 1881 Census Index by permission of the Latter Day Saints (LDS) church. The number of individuals in the 1881 Census was 30 million; in the NBI 6.4 million. The number of unique surname spellings in them was 404,793 and 175,688 respectively. SA had used the LDS table to encode the surnames and had expected the NBI index to form a part of the LDS index: surprisingly, a huge number of unique surnames in the NBI didn't overlap with the names in the LDS Index.

3.2 Points were raised concerning related issues: DT said in the 16<sup>th</sup> and 17<sup>th</sup> centuries there simply weren't standard surname spellings; Stella Colwell (SC) said one needed a specialist knowledge of dialect to recognise some surname spelling variants and that regional diversity must be considered; Else Churchill (EC) expressed a worry as an end user; Peter Christian (PC) commented that the grouping of surnames was the first step in locating variant. This latter: this was agreed.

3.3 KS explained that UKDA had long experience of building thesauri, and that experience suggested 100% accuracy could never be attained. SA illustrated some NBI pie-charts and graphs, and his attempted groupings of names. He didn't always agree with the groupings chosen by the LDS: if variants were to be grouped according to a standard, how did one decide what should be included? Should Burkinshaw and Burtenshaw be grouped together or not? Perhaps the user should be allowed to see the list of variants.

3.4 PC suggested the use of a phonetic algorithm, which SA had used. Edward Carney's book was relevant here. Peter Seaman (PS) mentioned that he had been working on four separate hearth tax collections for one place within the years 1670-1674: the same person could appear four times with widely different spellings, which might each at first sight need to be grouped under different standard names.

3.5 KS asked about provenance dependence; the notion might need to be built into the Thesaurus. IG warned that copies made from originals would produce variants far removed from the originals. Another delegate said that

AGREED

errors had to be entertained as well, to enable searches.

#### 4 **Outline of proposed project**

- 4.1 IG talked about key issues, resource requirements, and chiefly the need to establish objectives. Could an effective algorithm be developed? LDS grouping had been a useful start, but there were too many flaws in it. Points already made had been valuable. An on-line service was wanted so that a list of names could be put into a search engine, and it would identify the surname group to which each name belonged. Developing algorithms must bring in a knowledge of surnames and phonetics. Skilled manual dexterity was important. And what financial resources were available? Would charges be levied from services? He wanted the people round the table to define the project's objectives and say what resources they could contribute.

#### 5 **Open discussion**

##### 5.1 [DT said that the project should be focussed on what the user wanted.](#)

- 5.1 KS said why just apply one algorithm? Why not ten different ones? The more flexibility that was built into the system, the more people would believe it's a product that's worthwhile.
- 5.2 PC said the British Surnames Thesaurus Project should be regarded as an *improveable* project: scholars using it could pass on expertise and advise corrections.
- 5.3 David Crook (DC) said one controlling editorial mind was needed, not a committee: a project manager, with an IT expert, but not the same person.
- 5.4 *Copyright*. This was recognised by all to be an issue of importance.
- IG felt that if data was extracted (from a database or copyrighted source) and absorbed into a thesaurus and the public couldn't extract the original, then there wasn't a copyright problem. Peter Ruthven-Murray (PR-M) suggested that all holders of copyright might be persuaded of the greater public good. A discussion ensued on copyright.
- 5.5 *Objectives*. PR-M said a fundamental decision needed to be made whether to begin with a narrow or wide objective. If the British Surnames Thesaurus Project were to have an initial narrow objective, what user would searchers make of it? IG said it might be worth starting with a narrow objective, but with inbuilt flexibility. PC said this was an opportunity to start afresh, not to copy mistakes made already. SA mentioned a Mormon CD called *English Surname Groupings*.
- 5.6 IG reiterated an earlier point made by DT: concentrate first on user needs and benefits. John Challis (JC) asked, was this a service database? IG opined that it should be; the service would initially have only Anglicised versions of names. DC said that perhaps the starting date previously mentioned (of 1500) should be pushed back; PC said start at 1538. SC said the University of Leicester should be consulted on early modern British surnames.

- 5.7 *Resources.* KS said that UKDA could provide the project with space on a server, maintenance, technical support and staff to feed new data into the system, but that experts would be required to design the structure and architecture of the system. Similarly, [UKDA staff could implement the KS-could-provide](#) algorithms, but needed expert advice to [develop form](#) them. An editorial head and director were needed. KS wondered if the Marc Fitch Fund be approached?
- 5.8 DC asked if KS could he provide accommodation. KS replied that accommodation was a recurrent issue in all institutions if higher education, but KS could put it on the agenda. DT said that if funds were to be applied for, the University of Essex would the institution role very well. Else Churchill asked whether this had been done before and if we were actually reinventing the wheel. Louise Craven (LC) said she knew only of work done at the Getty to allow searching on linguistic variables of surnames, but not to enable searching of variants of spelling and variant names in the same language.
- 5.9 *Charging for information.* PC felt that the Thesaurus should be accessible for anyone to use: the dataset could be given out for anyone to work on, and a service should be provided to analyse new sets of data. Volunteer labour would have to be depended upon. Was this to be a charitable project, or would people have to pay for their profit-making? KS said it should not be for profit as the UKDA was funded to provide information to higher education bodies free of charge. The project could be extended to non profit-making historians, but had to get cost recovery from anyone else. IG said the emphasis should be on cost recovery, not money-making. It was agreed that the important thing was to get funds; a good project board was needed.
- 5.10 *Approach.* PR-M wondered whether an academic editorial approach would be used, or the rough-and-ready way of taking an initial thesaurus and working on the principle that anyone researching their name would know about it, therefore inviting their comment and tapping into their knowledge; deliberately to offer a flawed product, asking for improvements. PC said this already existed: the Guild of One-Name Studies. IG said there were drawbacks to this approach: it would mean a lack of control over provenance-related data. It would be very big and maybe full of rubbish; if the one-name method was followed, multiple groups would result.
- 5.11 PR-M said one issue to be taken into consideration when discussing objectives was whether we want a product we could use in one year's time or ten years' time? PC suggested starting with SA's NBI and the 1881 Census, as soon as possible. Documentation should be provided; some might not be available. SA had more than 100 rules for investigation. Some analysis had already been carried out by linguists; but then surname experts might further refine the data. SA had used Soundex for phonetic help.
- 5.12 JC said the 'name in a bucket' approach was flawed. A spelling should be given a weighting to refer it to the various standardisations. KS said that's what he meant by not having just one algorithm; have several. JC agreed, and suggested that rather than an editorial team, an all-automated programme to which users could contribute might be the way to go. This

AGREED

was not agreed to.

## **6 Proposal for Steering Group**

6.1 IG looked for volunteers.

6.2 David Thomas said the first step should be to write out a Project Initiation Document, which would set out all these issues. He volunteered to provide a skeleton for this; PC was willing to contribute.

6.3 IG said the Steering Committee should represent different interests. Nominated members of the Steering Committee were (in alphabetical order): Steve Archer; Peter Christian; Louise Craven; John Dawson; Ian Galbraith; Ben Laurie; Kevin Schurer; David Thomas.

6.4 The Project Initiation Document would be distributed to everyone attending the present meeting. KS agreed to set up a project website, which would contain the minutes of the meeting and any documents produced. IG said the [project](#) website should be closed and have a password, to restrict access, [though part might be open](#). The first meeting of the Steering Group would decide more about this.

6.5 It was decided that the first meeting of the Steering Group would take place at the Public Record Office on Wednesday 6 February 2002, at 2.00 pm. All present would be on the mailing list. Whether the entire company should meet again was a matter for further consideration.

## **7 Meeting close**

7.1 The meeting closed at 1.00 pm.

Peter Seaman  
11 December 2001