Enroller: An Experiment in Aggregating Resources

Jean Anderson

This chapter describes a collaborative project between e-scientists and humanists working to create an online repository of linguistic data sets and tools. Corpora, dictionaries, and a thesaurus are brought together to enable a new method of research. It combines our most advanced knowledge in both computing and linguistic research techniques.

Keywords: online repository, corpora, dictionaries, thesaurus, *Enroller*, interoperability, collaboration

Language scholars make use of a variety of language resources to conduct their research. Such resources include dictionaries, thesauri, corpora, audio and video collections. At present most of these resources are distributed, non-interoperable, and licence-protected. As a result researchers typically conduct their research through direct access to independent data sets using multiple browser windows and multiple authorisations. This approach results in non-scalable and less productive research, and can lead to incomplete or non-verifiable results. The JISC-funded project, *An Enhanced Repository for Language and Literature Researchers* (*Enroller*) addressed these issues through development of a targeted web-based research environment.[1] The project ran from 2009 to 2011 and continues to add data sets and enable corpus-based research.

Humanities scholars are creating a growing amount of digital data. We are told that Arts and Humanities Research Council funded research projects which produce some kind of digital output account for half their funding budget.[2] Digital data relevant to linguistic research are also produced outside Higher Education, for example dictionaries published by commercial companies and individual scholars' collections of texts, often from specific geographical areas. This is also the case for Scots, Scottish, and English materials. The *Enroller* project hypothesis was that bringing related data sets together in one online resource gateway – aggregating them – brings major benefits to researchers and makes better use of the national funding which helps us to create data sets.

Challenges

There are challenges in using digital data for research. Reports have shown that awareness and use of digital resources in the arts and humanities is low,[3] but opinions sought during conferences and workshops show that there is a great interest in learning more about advanced computing technologies and in sharing resources. Our participation in conferences and in meetings such as the SCOTS Symposium and the *Enroller* Colloquia in Glasgow has enabled us to identify data sets and functional requirements of interest to our linguistic colleagues and the challenges involved for them.[4]

Multiple identities

Researchers typically conduct their research through direct access to independent data sets using multiple browser windows and multiple authorisations requiring multiple identifiers and passwords to be remembered. At a practical level it is tedious and it can be discouraging to have to remember different passwords and to register to obtain permissions for many resources. We need methods of logging on and being recognised as bona fide researchers with one permanent (or at least long-lasting) identifier and password. The *Enroller* e-science team members endeavoured to find a solution to this problem.

Hidden data

Many data sets are held by individual scholars and are inaccessible to the community. We need to create systems which will allow the easy uploading of data to online platforms by resource creators who do not have the necessary programming skills or computing infrastructure to publish their data themselves. Data can be given to the Oxford Text Archive, but the OTA provides a limited search facility and does not have easily useable online access.[5]

Many data sets are deposited in national archives such as the OTA or local archives but are not known and therefore not used by the community. We need to create resource gateways which contain data sets and functions that scholars in particular fields want and that are easy for them to find. If we identify an area such as, for example, the English of northern Britain (Scots, Scottish English, Northumbrian English, etc.) we can liaise with scholars in the field and create networks which can allow them to describe their requirements, contribute their data, and test the gateways. There will be overlaps among disciplines and areas of study, and this could lead to further knowledge- and data-sharing.

Cross-searching

Currently we look up words in dictionaries, in thesauri, and in texts – one data set at a time. We cannot search several data sets at once. A researcher might want to look up a word such as *bonny* in the dictionary to find its meaning, in a thesaurus to look up the concepts and categories in which it is found, and in a corpus to find the documents containing it and the contexts of its use.[6] He or she might want to look at all the occurrences of a word in several corpora, limited by metadata such as geographical location, gender, or age of the writer or speaker. The researcher might also want to see a concordance and word frequency of the word in each found document. The user might want to save the results, make comparisons between them, and perhaps share them with other scholars. We currently have no available online facilities to do this.

Interoperability

Another challenge is that there is no interoperability between online resources. The lack of a method of communication between systems means that search results from one resource cannot easily be fed into queries to search another resource. For example, a researcher might want to look up a word in a dictionary and then see examples of use from a text archive or a corpus. This would be difficult to accomplish under current conditions, requiring the researcher to save and retain the results data from each search and try to find a way of using it with the next data set.

There are many reasons for the lack of interoperability: most of the problems are in the different formats that resource creators use. Standards exist but not every resource creator uses them, sometimes preferring for reasons of time or simplicity to use what they know. This results in resources having differing data formats, differing file-naming conventions, differing text encoding mark-up, and differing tools for analysis. XML is now the accepted standard meta-language for encoding text and should be the normal format to use in all textual projects.[7] The Text Encoding Initiative Guidelines advise upon the digital encoding of literary and linguistic texts.[8] (But even if TEI Guidelines are followed and XML mark-up chosen, there is a necessary flexibility which means that creators will apply the mark-up in different ways, adopting different sections of the Guidelines.) We cannot enforce one way of creating resources but we can encourage the use of methods which will alleviate the problem. We need a simple way of making this not just possible but easy for scholars.[9]

This situation becomes even more challenging when multiple data sets – dictionaries, thesauri, and text corpora – need to be cross-searched simultaneously. A lexicographer might want to consult the *Oxford English Dictionar*y, the *Scottish National Dictionary*, and the *Dictionary of the Older Scottish Tongue* to check definitions or find the earliest uses of a word, and then use the *Scottish Corpus of Texts & Speech*, the *Newcastle Electronic Corpus of Tyneside English*, and the *British National Corpus* to find examples of the word's use in modern contexts. Researchers will want to use the standard text analysis tools: concordances, word frequencies, and collocations. They want to save and download the results for further analysis or use targeted tools to investigate their hypotheses.

Concept searching

A humanities researcher might want to investigate a concept in use either currently or historically. It is easy to search for a word or words if you know them and their variant spellings. But it is difficult to know all the words that have been used to embody a particular concept through time and in different geographical locations. In dealing with non-standard language, historical or dialectal, we need to know we are finding all the references we need to draw reliable conclusions. For example, between AD 700 and the present there are nearly two hundred words in the category *woman* in the *Historical Thesaurus of English* (HTE). If a researcher today searched for *woman* in historical English texts it is probable that many relevant words would not be found. Here are some of them (the words are not all synonyms as pejorative terms are included):

wen, fæmne, freo, frowe, husbonde, ides, mæg_, meowle, virago, quean, wife, woman, lady, bride, carline/-ing, mare, female, stot, pigsn(e)y, piece, wye, fair, feminine, teg, she, minikin, pigsy, ware, jade, skirt, mort, mot, feme, pinnace, jug, tarleather, goddess, pussy, rib, sister woman, covess, pintail, wife and mother, buer, piece of goods, judy, femme, bit of muslin, shickster, fellow, Jack in petticoats, bint, popsy, tart, dona, ladykind, totty, tootsy, she-male, dame, frail, bit of stuff, floozie/floosie/floozy, muff, babe, bird …[10]

Having found the list of words that embody the relevant concept, the researcher might want to find all of the matching words in corpus resources and find out how the words have been used in different texts or at different times by different individuals, or how terms have been used in specific regions historically. The computational task becomes very demanding if the researcher decides to search for multiple, possibly hundreds, of words at once in several corpora and do all of the mentioned tasks simultaneously.

Most language and literature data environments do not permit scholars to do this; instead researchers are left with individual-level data sets, coded differently (e.g. with different metadata and data formatting), accessible through individual web-based interfaces with individual access codes and sometimes passwords. Most humanities departments do not have access to computational power sufficient to carry out the most intensive of these tasks.

Concept searching is currently an area of much research focus. The function would be of great value to commercial search engines systems like *Google*, *Yahoo!*, and *Bing*. It would be a benefit to the users of online journals, or indeed to the users of any system that searches text. The restriction of searching only for known words is an unnecessary limitation. Only practical considerations are in the way. Many computing science projects in this area are using thesauri to look up synonyms to use as search lists but the thesauri are not like the HTE in its size and authority.

The *Enroller* project

The above issues were addressed by the *Enroller* project, funded by the Joint Information Systems Committee (JISC) of the UK Government from 2009 to 2011. The aim of the project was to create a virtual research environment (VRE) based upon advanced e-science technologies and to test the use of the e-science Grid with humanities data and humanist research methods. The *Enroller* portal is now available.[11]

We envisaged it serving a broad spectrum of research on the languages and literature of Scotland, with the longer term intent of providing a model which could be extended to other languages and literatures. The partners in the venture were the *Software for Teaching English Language and Literature and its Assessment* project (STELLA) at the University of Glasgow,[12] the *National e-Science Centre* (NeSC),[13] *Scottish Language Dictionaries Ltd* (SLD),[14] and the *Newcastle Electronic Corpus of Tyneside English* (NECTE).[15]

The background

The project required collaboration between computer scientists and scholars of language and literature to achieve its aims and address the problems and questions above. The University of Glasgow was well placed to bring together these communities, with many years of experience in creating and directing digital projects for the humanities and also hosting NeSC.

The School of Critical Studies includes the departments of English Language, English Literature, and Scottish Literature who have been leaders in the field of literary and linguistic computing since the creation of the STELLA project in 1987. Members of the School have created ongoing digital resources (corpora, thesauri, hypertext editions, computer-based teaching programs) and engage in digitally-based research.[16] The School has the only department of Scottish Literature in the world and the English departments involved in the project contain internationally-known linguistic scholars. It also has access to data sets and expertise in Scots, a complex and diverse linguistic variety, which made it an ideal test-bed for the project. The humanities members of the *Enroller* team were Jean Anderson as

Principal Investigator, Marc Alexander as Research Associate, and Johanna Green as Research Assistant.

NeSC has a research portfolio across a broad range of areas and experience in relevant technologies including portal technologies, web service technologies, and security technologies. The NeSC *Enroller* team members were Professor Richard Sinnott as Co-investigator and Mohammad Sarwar as Programmer. Dr John Watt took over from Professor Sinnott when the latter left the University.

The staff of NeSC use *ScotGrid*, the Scottish Higher Educational Funding Council funded high-powered computing cluster for the analysis of data. *ScotGrid* is primarily used by physicists working on particle detector experiments at the Large Hadron Collider – it is rarely used by humanists. Grid computing is using many computers in a network on a single problem at the same time. Together, many ordinary computers have the power of a supercomputer. A grid needs special software to parcel out the work and put the results back together for the user. There are four Grids in the UK: *ScotGrid*, *NorthGrid*, *LondonGrid*, and *SouthGrid*.[17] The Science and Technology Facilities Council described the Grid as follows:

> The Grid is the next step in exploiting networked computer power. Currently the Internet and World Wide Web allow us to share information and transfer data quickly and easily around the world. In the future the Grid will let us share computer processing power, software packages and data storage space. The Grid has many applications, but its first major application will be to allow researchers at CERN to share global computing power to manage and process the huge quantities of data that will be produced by the LHC. By linking desk top computers in a global network, managed by 'middleware', the Grid brings supercomputing power to desk tops.[18]

For more information on the Grids see Professor Sinnott's presentation,[19] the Arts and Humanities e-Science Support Centre and *ScotGrid*'s web pages.

Humanities research does not normally require such high-powered computing as even the largest texts are very much smaller than the numerical data files produced by science experiments. However, we thought that the comparatively complex structure of language combined with complicated searches over many data sets would require much more computing power than is usually available to humanities scholars. This was an exciting new venture and both humanists and e-scientists looked forward to discoveries and transfer of knowledge between widely separated disciplines.

The intended benefits to linguists and to computing scientists

Enroller was intended to give researchers in Scottish and English languages and literature access to large amounts of data from a single, easy-to-use portal. They would have membership of an international network of scholars. We hoped to disseminate increased knowledge of digital resources and standard formats and methods. The portal would both provide wider access to the resources and raise awareness and understanding of e-science. It would bring together resources and tools for scholars who were already familiar with ICT in research, and introduce new users to advanced computing tools and methods, as well as encouraging increased use of existing electronic resources. It would allow a community of researchers with related aims to collaborate more easily, and already-funded data sets to be used in new combinations that could result in heuristic discoveries. The wider humanities community would benefit from the models developed here. The resulting knowledge transfer would be of benefit to both the humanities and the e-science communities as well as to the wider community such as publishers, dictionary creators, and national services.

For the e-scientists in NeSC there were numerous issues to overcome to deliver Grid solutions within this unfamiliar context. Examples of these challenges included ensuring the provenance of data and its authenticity; dealing with the intellectual property and copyright associated with the differing distributed data sets; dealing with heterogeneous data resources and different data formats; providing a range of integrated tools that researchers can use for finding, querying, and storing data; and designing an interface to present the results of searches performed on many data sets. These functions need to be presented in such a way that users are untroubled by the fact that they are undertaking their research ‘on the Grid’. For the most part, the Grid community has not yet fully achieved the true vision and potential of the Grid and has concentrated on developing middleware software rather than addressing research needs. It was also thought that the language and literature community and other non-scientists would not adopt science Grid models which require the users to register for digital certificates every time they want to use the system, to submit jobs, and wait, perhaps for a few days, to receive the search results. This assumption was found to be true during the tests of the initial *Enroller* prototype systems by the Network of Scholars.

Instead these users demand seamless and intuitive e-infrastructure supporting their daily research and this is what we aimed to provide.

The data resources brought together in *Enroller*

Data sets were contributed by scholars from Glasgow, Edinburgh, and Newcastle. These are essential research tools for anyone interested in the languages and dialects of northern Britain, and for historical or literary scholars whose sources are written in these varieties of English.

The *Scottish Corpus of Text & Speech* (SCOTS) was funded by the EPSRC and the AHRC and developed in the University of Glasgow from 2001 to 2007.[20] It is a collection of text, video, and audio files covering the period from 1945 to the present. It has 4.5 million words, including transcriptions of the audio and video materials. The texts are in Scots, Scottish English, and Standard English. *Enroller* used only the textual data as the project did not allow time for the development of search and presentation facilities for audio and video material. Also the project did not intend to duplicate the excellent interface and functions already available on the SCOTS web site.

The *Newcastle Electronic Corpus of Tyneside English* (NECTE) is an AHRC-funded, University of Newcastle corpus of dialect speech from Tyneside in Northeast England.[21] It is based on two corpora, one of which was collected in the late 1960s by the Tyneside Linguistic Survey (TLS) project and the other which was created in 1994 by the Phonological Variation and Change in Contemporary Spoken English (PVC) project.[22] NECTE has brought the TLS and PVC materials together in a single corpus and makes them available in a variety of formats: digitised audio, standard orthographic transcription, phonetic transcription, and part-of-speech tagged. These parts of the corpus are aligned so that the user can, for instance, read the transcription of a conversation while hearing the audio or see the orthographic and phonetic transcriptions side by side. The NECTE corpus is tagged in Text Encoding Initiative conformant XML. Again only the textual data is included in *Enroller* as the funding did not allow time to implement new parts of the system to cater for audio.

The *Dictionary of the Scots Language* (DSL) is an online resource created by Scottish Language Dictionaries.[23] It contains the two main dictionaries of Scots, the *Dictionary of the Older Scottish Tongue* and the *Scottish National Dictionary*.[24] Together they give a comprehensive history of Scots language and culture over the last eight hundred years. The *Dictionary of the Older Scottish Tongue* covers the period of Scots from the earliest records in the twelfth century to about 1700, and the *Scottish National Dictionary* covers the period from 1700 to the latter half of the twentieth century. It is updated with supplements, the most recent of which was added in 2005. Each entry shows the chronological and semantic development of a Scots word and gives details of orthographic variants, grammatical inflections, derivative words and phrases, and etymological history. The definitions are illustrated by quotations from over six thousand sources, covering a wide range of subject areas. Many of the modern Scots words are also supported by evidence from oral sources, and include information on phonological and dialectal variation.

The *Historical Thesaurus of English* (HTE) is the world's largest thesaurus and the most complete thesaurus of English.[25] It arranges all the recorded meanings in English from Anglo-Saxon times to the present into hierarchical semantic categories. It is the only historical thesaurus of English, was created by internationally renowned lexicographers, and has a new, expertly devised classification system. A new ontology was needed because Roget's classifications were not detailed enough for linguistic research and were idiosyncratic, being created by one man in the early nineteenth century. Most available thesauri and semantic taggers are based on Roget's outdated classifications, including WordNet and the Lancaster UCREL Semantic Analysis System (USAS) commonly used by computer scientists.[26] These semantic taggers are unfortunately not appropriate for humanities research which requires detailed and fine-grained results, or which work on complex or specialist or historical texts. The USAS tagset has 235 categories, one thousandth of that available in the HTE, and WordNet has 117,659 'synsets' (semantic categories). HTE has a level of detail unseen before now: 800,000 meanings for 600,000 words in over 235,000 categories; a detailed, numerically-tagged semantic hierarchy; a comprehensive index enabling complete cross-referencing; synonyms listed with dates of first recorded use in English, in chronological order. On its print publication by Oxford University Press in 2009,[27] HTE was greeted with international acclaim as 'a monumental work of scholarship' (Professor David Crystal); 'the thesaurus by which all others are judged' (author Phillip Pullman); 'Forty-five years of exacting scholarship by a well-led team have had a triumphal outcome. This book is a magnificent achievement of quite extraordinary value. It is perhaps the single most significant tool ever devised for investigating semantic, social, and intellectual history.' (Professor Randolph Quirk). The Thesaurus is now used as part of the *Online Oxford English Dictionary* to provide synonyms for the words searched for.[28] HTE is held in a digital database with content improved and extended since the print publication. The inclusion of the HTE in *Enroller* allows one of the most powerful functions to be implemented, concept searching.

Since the active funding phase of the project ended we have added a few more data sets to *Enroller*.

The *Helsinki Corpus of English Texts* is a diachronic corpus, which includes text samples from Old, Middle, and Early Modern English.[29] The texts are organised by period and each one has a list of codes giving information on the text and its author. In *Enroller* we have split the corpus into its three component parts and have one corpus for each period. We are grateful to the Department of English, University of Helsinki, for permission to include these three prestigious collections. The corpus was previously available only on CD-ROM and was encoded in the COCOA format used by the now superseded *Oxford Concordance Program* (OCP),[30] and *Text Analysis Computing Tools* (TACT).[31]

Hansard 1805-2005 is a 2.3 billion word corpus containing texts of speeches in both the House of Lords and the House of Commons, and the Committees of Parliament. The addition of this huge data set to *Enroller* was funded by JISC in the University of Glasgow *Parliamentary Discourse* project in late 2011.[32] The Hansard records of debates in Parliament from 1803 to 2005 have been digitised and encoded in XML format by the libraries of the House of Commons and the House of Lords, for parliamentary use. The data had not been annotated for any further uses. This has provided academic and non-academic communities with a significant amount of high-quality data but with limited use. The data is available under an Office of Public Sector Information licence from Parliamentary Services.[33] These debate records are already of significant usefulness for scholars of British politics and history but the data is limited in its exploitation by linguists as it is unannotated for anything beyond simple text searches.

The *Parliamentary Discourse* project took this primary source and enhanced it by adding linguistic annotations, topic mark-up, and named entity tags, and by cross-linking it to the resources in the *Enroller* portal such as the HTE. This enables searching of textual resources by concepts and across time. Researchers can therefore investigate not only who said what and when, but also find all the references to a topic even when not all the possible words were known to the user.

Network of Scholars

Enroller has an associated Network of Scholars who contributed to the design and testing of the VRE and used the resources in research projects in their fields. The thirty-one scholars came from universities in the UK mainly, with a few colleagues from Europe. They participated by email using a project discussion list (JISCmail),[34] by attending two colloquia to discuss requirements, and by attending three workshops to test the system.

The portal

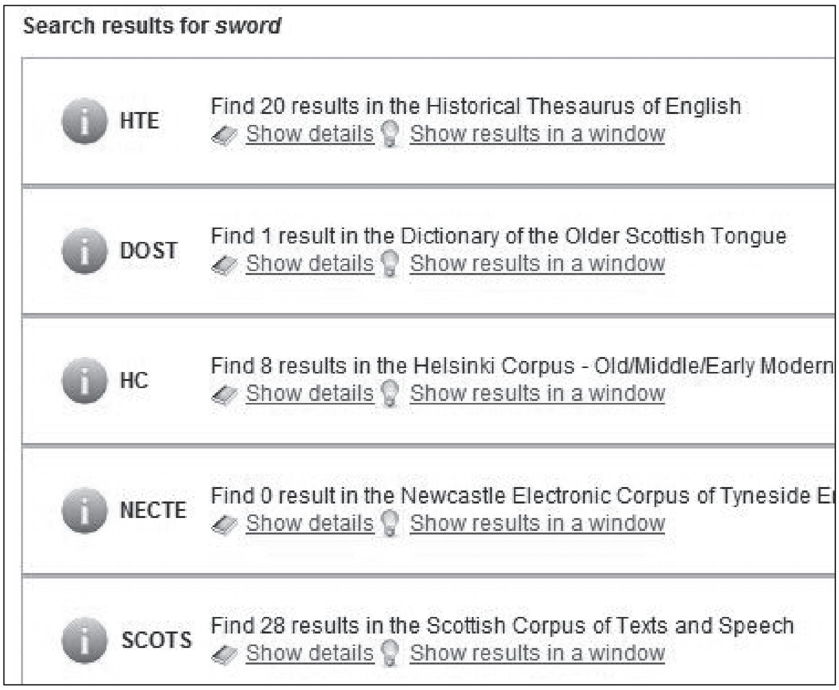
The VRE consists of a web portal giving access to several data sets. In response to the requirements described above the portal was designed to provide interoperability between the online resources, having an easy-to-use search system, with tools for linguistic analysis, including concordancing and frequency of occurrence information where appropriate to the data. The system has support for security of access to licensed data by the authentication of users through the UK Access Management Federation.[35] This means that licences and copyright permissions which are agreed by educational establishments are applied to the individual user once logged in. At login the individual's own organisation's ID and password are used and no new IDs or passwords need to be created and remembered to use *Enroller*.

There are two types of search in *Enroller*. In the Basic Search the user can type in a word, phrase, or list of words and search one or more of the resources. The results can be copied to the user's own computer but not saved within the *Enroller* system. In the Advanced

Search the user can save results files and use them as input to further searches and also access *ScotGrid* to perform computationally intensive searches. An example workflow of the Advanced Search would be firstly to log in using the user's own university ID and password; search the HTE for all synonyms of the word *bonny*; save the results file; go to the Advanced Search menu; upload the saved text file; choose the corpora to be searched; submit the search to *ScotGrid*; access and save the results file when notified that it is ready.[36] Basic searches can be performed without authentification but to use the Advanced Search the user must first log in.

Cross searching

All of the data sets listed above can be searched separately or together. What follows are some examples of the results found when searching all the data sets simultaneously for the word *sword*.



Search results for *sword*

HTE
Find 20 results in the Historical Thesaurus of English
Show details Show results in a window

DOST
Find 1 result in the Dictionary of the Older Scottish Tongue
Show details Show results in a window

HC
Find 8 results in the Helsinki Corpus - Old/Middle/Early Modern
Show details Show results in a window

NECTE
Find 0 result in the Newcastle Electronic Corpus of Tyneside E
Show details Show results in a window

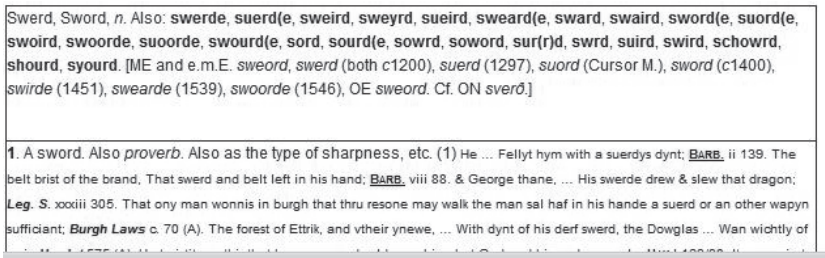
SCOTS
Find 28 results in the Scottish Corpus of Texts and Speech
Show details Show results in a window

Figure 1: Example results of search for *sword* in the data sets

First a table is shown giving the number of results for each data set. For each set the user can choose to see more details, to save the results to their own computer for later analysis, or to compare two results side by side.

Figure 1 shows part of the initial results table after all data sets have been searched for the word *sword*. Clicking on the 'Show details' link will display a list of all the entries found. Clicking on an item in the list will display the full text of that item in a small window. Clicking the 'Select for comparison' button will mark the item for display with other marked items in side-by-side windows for ease of reading. There follow some examples of actual data found when searched for *sword*.

Figure 2 shows part of the text window displayed when the DOST 'Show details' is clicked. The whole dictionary entry is available including the headword, variants, and citations.



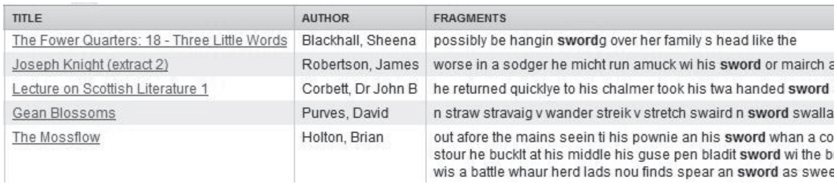
Swerd, Sword, *n.* Also: **swerde**, **suerd(e**, **sweird**, **sweyrd**, **sueird**, **sweard(e**, **sward**, **swaird**, **sword(e**, **suord(e**, **swoird**, **swoorde**, **suoorde**, **swourd(e**, **sord**, **sourd(e**, **sowrd**, **soword**, **sur(r)d**, **swrd**, **suird**, **swird**, **schowrd**, **shourd**, **syourd**. [ME and e.m.E. *sweord*, *swerd* (both c1200), *suerd* (1297), *suord* (Cursor M.), *sword* (c1400), *swirde* (1451), *swearde* (1539), *swoorde* (1546), OE *sweord*. Cf. ON *sverð*.]

1. A sword. Also *proverb*. Also as the type of sharpness, etc. (1) He … Fellyt hym with a suerdys dynt; BARB. ii 139. The belt brist of the brand, That swerd and belt left in his hand; BARB. viii 88. & George thane, … His swerde drew & slew that dragon; ***Leg. S.*** xxxiii 305. That ony man wonnis in burgh that thru resone may walk the man sal haf in his hande a suerd or an other wapyn sufficiant; ***Burgh Laws*** c. 70 (A). The forest of Ettrik, and vtheir ynewe, … With dynt of his derf swerd, the Dowglas … Wan wichtly of

Figure 2: DOST: the first part of the entry for *sword*

This DOST example demonstrates a real challenge in dealing with non-standard texts; that of variant spellings. It is very difficult if not impossible to identify all the probable spellings of a word in historical times or in dialectal texts. This problem is not addressed in the *Enroller* project but it is an important one that deserves effort and funding to ameliorate.

Figure 3 shows part of the results for SCOTS displayed when 'Show details' is clicked. The display gives the title, author, and lines of text where the word occurs with the word highlighted. The list can be easily sorted alphabetically by clicking on 'Title', 'Author', or 'Fragment'. The 'Fragment' section is a concordance display which allows the user to quickly see the differing contexts in which the word is used.

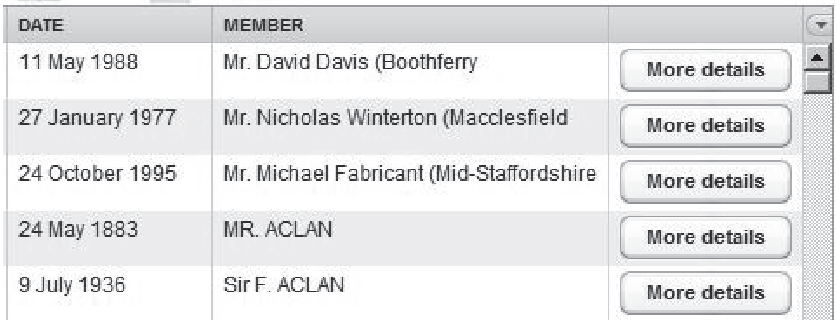


TITLE	AUTHOR	FRAGMENTS
The Fower Quarters: 18 - Three Little Words	Blackhall, Sheena	possibly be hangin **sword**g over her family s head like the
Joseph Knight (extract 2)	Robertson, James	worse in a sodger he micht run amuck wi his **sword** or mairch a
Lecture on Scottish Literature 1	Corbett, Dr John B	he returned quicklye to his chalmer took his twa handed **sword**
Gean Blossoms	Purves, David	n straw stravaig v wander streik v stretch swaird n **sword** swalla
The Mossflow	Holton, Brian	out afore the mains seein ti his pownie an his **sword** whan a co stour he bucklt at his middle his guse pen bladit **sword** wi the b wis a battle whaur herd lads nou finds spear an **sword** as swee

Figure 3: SCOTS: five of twenty-eight results

Clicking on the underlined title will display a text window with the whole text of the named document. Often a researcher will want to see more context than the concordance allows before making an analysis of the word in the particular instance of use.

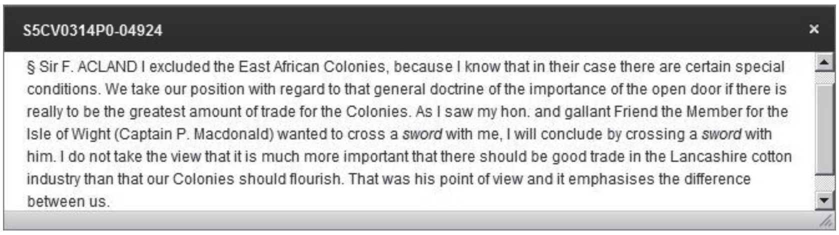
Figure 4 shows part of the list of speeches made in the House of Commons containing the word *sword*. The list could be sorted by date or by Member of Parliament's name.



DATE	MEMBER	
11 May 1988	Mr. David Davis (Boothferry	More details
27 January 1977	Mr. Nicholas Winterton (Macclesfield	More details
24 October 1995	Mr. Michael Fabricant (Mid-Staffordshire	More details
24 May 1883	MR. ACLAN	More details
9 July 1936	Sir F. ACLAN	More details

Figure 4: Commons: the first five of 2172 results of a search for *sword*

Clicking the 'More details' button will display the whole speech in a scrollable text window with the search word highlighted. The researcher can read the full context of the word's use.



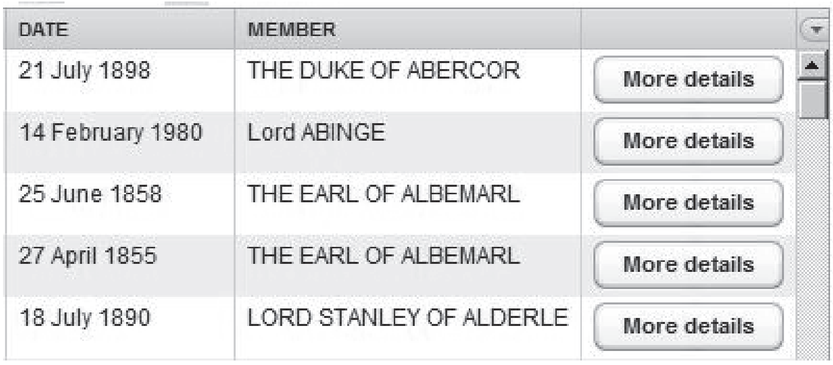
S5CV0314P0-04924

§ Sir F. ACLAND I excluded the East African Colonies, because I know that in their case there are certain special conditions. We take our position with regard to that general doctrine of the importance of the open door if there is really to be the greatest amount of trade for the Colonies. As I saw my hon. and gallant Friend the Member for the Isle of Wight (Captain P. Macdonald) wanted to cross a *sword* with me, I will conclude by crossing a *sword* with him. I do not take the view that it is much more important that there should be good trade in the Lancashire cotton industry than that our Colonies should flourish. That was his point of view and it emphasises the difference between us.

Figure 5: Extract of the speech by Sir Francis Acland made in July 1936

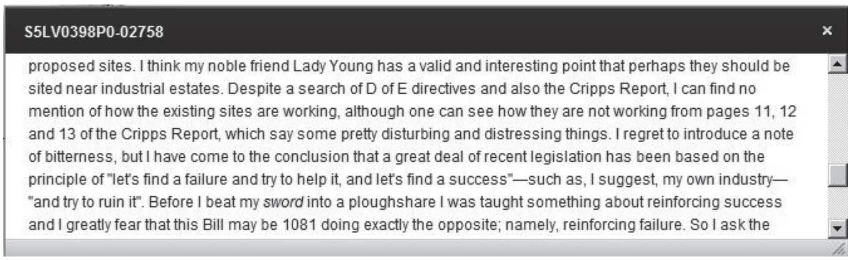
Sir Francis Dyke Acland, 14th Baronet, was a British Liberal politician. He served as Under-Secretary of State for Foreign Affairs between 1911 and 1915. Note that both times he is mentioned (the first is before he acceded to the peerage in 1926) his name is wrongly given as 'ACLAN' in the Parliamentary archive and not 'Acland'. It is not unusual for there to be mistakes in data, especially in old data. I have not investigated whether the mistake is in the first, printed edition of Hansard or was introduced during digitisation. The possibility of mistakes in data should always be remembered by researchers. It is their responsibility to ensure correct results and not to rely on the computer always being right!

Figure 6 shows the list of results from speeches in the House of Lords. Again the list could be sorted by date or by Member's name and asking for 'More details' displays the whole speech with search word highlighted in italics (Figure 7).



DATE	MEMBER	
21 July 1898	THE DUKE OF ABERCOR	More details
14 February 1980	Lord ABINGE	More details
25 June 1858	THE EARL OF ALBEMARL	More details
27 April 1855	THE EARL OF ALBEMARL	More details
18 July 1890	LORD STANLEY OF ALDERLE	More details

Figure 6: Lords: a selection of 610 results



S5LV0398P0-02758 ×

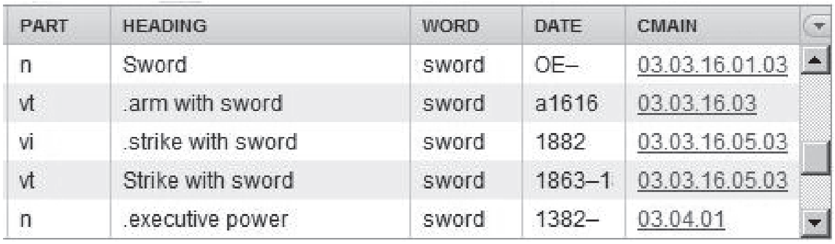
proposed sites. I think my noble friend Lady Young has a valid and interesting point that perhaps they should be sited near industrial estates. Despite a search of D of E directives and also the Cripps Report, I can find no mention of how the existing sites are working, although one can see how they are not working from pages 11, 12 and 13 of the Cripps Report, which say some pretty disturbing and distressing things. I regret to introduce a note of bitterness, but I have come to the conclusion that a great deal of recent legislation has been based on the principle of "let's find a failure and try to help it, and let's find a success"—such as, I suggest, my own industry—"and try to ruin it". Before I beat my *sword* into a ploughshare I was taught something about reinforcing success and I greatly fear that this Bill may be 1081 doing exactly the opposite; namely, reinforcing failure. So I ask the

Figure 7: Extract of the speech by Lord Stanley of Alderley made in July 1890

With this range of results a researcher can examine definitions and uses of a word in a variety of contexts and over a long period of time.

Concept searching

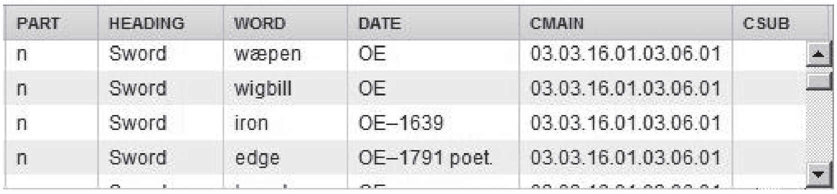
The above examples show the results of a simple word search. If as mentioned above the user is interested in a concept and *all* the words that have been used to embody it in English then a concept search is available by saving the HTE list of synonyms and using that list to search the chosen resources.



PART	HEADING	WORD	DATE	CMAIN
n	Sword	sword	OE–	03.03.16.01.03
vt	.arm with sword	sword	a1616	03.03.16.03
vi	.strike with sword	sword	1882	03.03.16.05.03
vt	Strike with sword	sword	1863–1	03.03.16.05.03
n	.executive power	sword	1382–	03.04.01

Figure 8: HTE: *sword* occurs under several category headings (meanings)

Choosing the first entry in this list will display all of the words in this category in chronological order of first attestation:



PART	HEADING	WORD	DATE	CMAIN	CSUB
n	Sword	wæpen	OE	03.03.16.01.03.06.01	
n	Sword	wigbill	OE	03.03.16.01.03.06.01	
n	Sword	iron	OE–1639	03.03.16.01.03.06.01	
n	Sword	edge	OE–1791 poet.	03.03.16.01.03.06.01	

Figure 9: The first part of the list of synonyms for *sword* in the HTE

The list goes on to show over two hundred words embodying the meaning of some kind of sword:

… wæpen, wigbill, iron, edge, brand, sword, brandellet, gare, tool, brank(e), tranchefer, flatchet, morglay, smiter, toasting-iron, brandiron, toledo, spit-frog, spit, spurtle, bilbo, tilter, porker, degen, slasher, cheese-toaster, toasting-fork, windlestraw, brad sweord, bill, falchion, glaive, broadsword, claymore, scimitar, shamsheer, badelar, sabre, rapier, bird-spit, walking-rapier, walking-sword, single-rapier, small-sword, epée, …

This is about half of the actual number of words under the ‘sword’ category in the HTE. It demonstrates clearly how a researcher could miss mentions of swords in text searches. *Enroller* provides a rare possibility to search for a concept because we have brought together the HTE and several data resources. The list of synonyms can be saved and used to search all the *Enroller* data sets by logging in and choosing the Advanced Search as described above.

The concept search is of interest to both linguists and historians. For example, colleagues in Glasgow working on the *Parliamentary Discourse* project will use this method to examine the concept of ‘integrity’ as referred to in the UK Parliament.[37] Linguists and historians will be able to link key words and concepts to historical and political events of their date of use.

The results of these large searches will contain many words that are not of interest to the researcher because of the large number of homographs in English, that is, words that are spelled the same but have different meanings. For example, *sewer* – ‘seamstress’ or ‘conduit for waste water’; *bear* – ‘animal’ or ‘to carry’; *wind* – ‘movement of air’ or ‘to meander’. If a researcher uses the list of synonyms for *sword* mentioned above to search texts for all the words which have been used to refer to a kind of sword, the results will include many unwanted cases as the list above contains several homographs: *iron*, *edge*, *brand*, *spit*, etc. This ‘noise’ has to be removed manually from the results before analysis. This is tedious but the work involved is compensated for by the richness of the relevant results that could only be otherwise found by weeks of manual trawling of texts.

Conclusion

The *Enroller* project proved that we could bring together several resources and successfully search them simultaneously. We have shown that interoperability is possible and we have generated much discussion about this means of research. We have presented papers at over a dozen conferences (both arts and e-science meetings) and been invited to give talks on the project at four international gatherings. Interest in the aggregation of resources is still clearly high and scholars are aware that there are still many challenges ahead and practical decisions to be made.

We found substantial challenges in implementation. The lack of use of well described XML encoding of texts meant that we had to spend more time than was anticipated in transforming texts to fit into the *Enroller* format. We soon discovered that dealing with many different formats was not a possibility as the amount of programming time required to create separate search systems for each different format was prohibitive and not sustainable.

The multiple-level display interface has proved to be one of the most difficult parts of the *Enroller* system to implement. The linguistics development team and the Network of Scholars all had different ideas of how best to present the complex results and several versions of the interface were developed before we settled on the current one. It has proved to be easily useable by our testers. The e-science implementation team learned how complex language really is and how disparate linguistic opinions can be. We have ended with a workable solution but we are aware that it is not perfect. Interfaces like this will be improved and the functions they make available will be taken forward in future projects.

Enroller has been described as a pioneer for the provision of digital resources for linguistics and as a leader in the community. We were asked to be founder members of ClariNet UK, a new organisation which will try to pull together all the disparate resources. This is an off-shoot of the European Union *Common Language Resources and Technology Infrastructure* project, CLARIN, which defines itself as:

> […] a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily useable for the whole European Humanities (and Social Sciences) community. CLARIN will offer scholars the tools to allow computer-aided language processing, addressing the multiple roles language plays in the Humanities and Social Sciences.
>
> The European Resources Infrastructure that is to be created by CLARIN is based on an open European Federation of strong service centres and repositories that jointly provide
>
> (i) knowledge about the existence of language resources,
>
> (ii) coordinated creation of, archiving of, and access to such resources,
>
> (iii) access to services and tools that would allow scholars to operate on such resources,
>
> (iv) bundling of and access to expertise related to specific language processing problems.[38]

It is important that linguists engage with these well-funded, international projects and ensure that linguists' and other humanists' requirements drive the design and implementation of the portals. The *Enroller* project has highlighted again the gap in the mutual knowledge and understanding of the methods used and goals sought by humanists and computer scientists. We expect the future for linguistics and all its branches to be one of aggregation and interoperability, and collaboration with technology.

[1] The *Enroller* portal can be accessed at <http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/enroller/ > [this and all other links in this chapter were accessed on 21 July 2012].

[2] For example, Claire Warwick, Melissa Terras, Paul Huntington, Nikoleta Pappa, and Isabel Galina, 'The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data' (University College London, 2006). Available online at <http://www.ucl.ac.uk/slais/claire-warwick/publications/LAIRAHreport.pdf>.

[3] Stuart Dunn, 'E-Science in the Humanities: Collaboration, Data and Processes' (unpublished conference paper at Digital Resources for the Humanities & Arts, Dartington, 2006). Conference programme available online at <http://projects.oucs.ox.ac.uk/DRHA/2006/drha2006-programme.pdf>.

[4] Details of the last SCOTS Symposium can be found at <http://www.scottishcorpus.ac.uk/about/symposium2006/>; details of the *Enroller* Colloquia are at <http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/enroller/colloquia/>.

[5] The UK government-funded OTA has, since the 1970s, accepted texts created by scholars and made them available freely to others. See <http://ota.ahds.ac.uk/>.

[6] *Bonny* is defined as 'beautiful, pretty, fair' in the *Dictionary of the Scots Language*, <http://www.dsl.ac.uk/dsl>.

[7] The Extensible Markup Language, XML, is described at <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html>.

[8] The Text Encoding Initiative, TEI, provides Guidelines for tagging text at <http://www.tei-c.org/index.xml>.

[9] For a full discussion on standards, see *A Companion to Digital Humanities*, eds. Susan Schreibman, Ray Siemens, and John Unsworth (Oxford: Blackwell, 2004), also available online at <http://www.digitalhumanities.org/companion/>.

[10] The *Historical Thesaurus of English* main web page is at <http://libra.englang.arts.gla.ac.uk/historicalthesaurus/>.

[11] The *Enroller* portal is at <https://enroller.nesc.gla.ac.uk/>.

[12] Information about STELLA can be found at <http://www.gla.ac.uk/schools/critical/aboutus/resources/stella/>.

[13] NeSC can be found at: <http://www.nesc.ac.uk/>.

[14] SLD is at <http://www.scotsdictionaries.org.uk/index.html>.

[15] NECTE is a project of the Department of English at the University of Newcastle: <http://www.ncl.ac.uk/necte/>.

[16] For more information on the School of Critical Studies and the STELLA project see <http://www.gla.ac.uk/departments/englishlanguage/research/> and <http://www.gla.ac.uk/schools/critical/aboutus/resources/stella/>.

[17] The UK Grids are described on the *ScotGrid* pages at: <http://www.scotgrid.ac.uk/>.

[18] Science and Technology Facilities Council, <http://www.lhc.ac.uk/about-the-lhc/what-is-the-lhc/11844.aspx>.

[19] Professor Sinnott gave a talk on the 'e-Context of *Enroller*' at the first *Enroller* Colloquium in Glasgow in 2009: see <http://www.slideshare.net/EnrollerProjectGlasgow/sinnott-paper>.

[20] SCOTS: <http://www.scottishcorpus.ac.uk>.

[21] NECTE: <http://www.ncl.ac.uk/necte>.

[22] More information on the Tyneside Linguistic Survey can be found at: <http://research.ncl.ac.uk/decte/tls.htm>. The Phonological Variation and Change in Contemporary Spoken English web page is at: <http://research.ncl.ac.uk/decte/pvc.htm>.

[23] The *Dictionary of the Scots Language* can be searched at: <http://www.dsl.ac.uk/dsl>.

[24] More information on the *Dictionary of the Older Scottish Tongue* is at: <http://www.scotsdictionaries.org.uk/Publications/DOST.html>; a description of the *Scottish National Dictionary* can be found on the Scottish Language Dictionaries web page at <http://www.scotsdictionaries.org.uk/Publications/SND.html>.

[25] The *Historical Thesaurus of English* main web page is at: <http://libra.englang.arts.gla.ac.uk/historicalthesaurus/>.

[26] WordNet is available at: <http://wordnet.princeton.edu/>, and Lancaster USAS at <http://ucrel.lancs.ac.uk/usas/>.

[27] *Historical Thesaurus of the Oxford English Dictionary*, ed. by Christian Kay, Jane Roberts, Michael Samuels, and Irené Wotherspoon (Oxford: Oxford University Press, 2009).

[28] The *Online Oxford English Dictionary* now includes the *Historical Thesaurus of English*. See <http://www.oed.com/>.

[29] The *Helsinki Corpora of English* are described at the University of Helsinki web page: <http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/>.

[30] OCP was the first and for many years the most successful text analysis program for use on IBM-type personal computers. It revolutionised linguistic research methods. It was developed by Susan Hockey and programmer Ian Marriott. See <http://digital.humanities.ox.ac.uk/about/history.aspx>.

[31] TACT is a multilingual text retrieval system which worked under the MS-DOS (pre-Windows) operating system on IBM-type PCs. It was developed at the University of Toronto by a collaborative team of textual scholars and programmers under the leadership of Ian Lancashire. TACT was very widely used because of its many parts which could be used for work from simple searches to complex statistical functions. See <http://projects.chass.utoronto.ca/tact/TACT/tact0.html>.

[32] The *Parliamentary Discourse* project is ongoing. More information can be found at: <http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/parliamentarydiscourse/>.

[33] Parliamentary Services is the UK Government body which manages current and archived data. Information on licensed use can be found at: <http://www.parliament.uk/business/publications/parliamentary-archives/archives-practical/copyright-and-use/>.

[34] The UK Joint Information Systems Committee provides a free email discussion service for academia. See <http://www.jiscmail.ac.uk/>.

[35] See <http://www.ukfederation.org.uk/content/Documents/HowItWorks>.

[36] This can take some time if the physicists are making heavy use of the Grid.

[37]See:<http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/parliamentarydiscourse/>.

[38] To find out more or to join CLARIN, go to: <http://www.clarin.eu/external/index.php>.