Technical Plan

Section 1: Summary of Digital Outputs and Digital Technologies

The project will produce a freely available online atlas of Scots syntactic features through which data such as audio recordings and geospatial, sociological and biographical material will be accessible. Two online atlas interfaces will be developed: one for academics that will feature in-depth search facilities and access to a broader selection of data fields, and one for a more general audience that will present users with streamlined search facilities, a less extensive set of data fields, and options to contribute to the project, for example by completing questionnaires. Both resources will be free to access, will be built upon a 'Google Maps' style interface and will feature extensive commentaries and summaries.

Users will be able to dynamically filter the survey locations that are displayed on the map based on criteria such as observed linguistic phenomena (e.g. progressive use of statives) and the biographical properties of the speakers (e.g. age, gender). Alternative visualisation layers such as heat maps, isoglosses and choropleth maps will also be available and will be generated dynamically from the underlying data based on a user's selected criteria. Through the map, users will also be able to click on survey locations to access the relevant data, such as audio recordings of informant interviews and their transcriptions, relevant judgement data for individual linguistic phenomena and anonymised biographical information.

In addition to the map-based means of accessing the data, users will also be able to search and browse the textual data and will be given the option of jumping from the textual view of their search results to the map view (and vice-versa) at the click of a button. An API (application programming interface) will also be developed for direct access to the project data, enabling future applications to query the data or for the entire dataset to be exported for future use.

In addition to the front-end, a database and online content management system (CMS) incorporating editorial workflows will also be developed for the project, through which the survey data (including textual, geospatial and audio data) will be uploaded, edited and published. The CMS will incorporate a version of the map-based interface that will enable researchers to associate survey data with a location simply by clicking on the map.

Both of the publicly available online interfaces and the CMS will be designed to be 'responsive', enabling their use on a variety of devices from desktop PCs to smartphones and tablets. This will be important not just for end users who are moving away from traditional means of accessing the internet but also for researchers who may be inputting data whilst engaging in fieldwork. An initial project website incorporating a blog will also be established at the start of the project and will be regularly updated.

Section 2: Technical Methodology

2a: Standards and Formats

All interviews will be recorded and stored in uncompressed WAV format (LPCM-encoded, two channel, 44100 samples per second, 16 bits per sample). The judgement elicitation interviews will last 60 to 90mins each, and the spontaneous conversation recordings will be 60mins long. At each of the 122 locations there will be four judgement elicitations, two conversation recordings and additional recordings of example sentence for use in interviews. This will amount to as much as 1000 hours of recordings, requiring about 600GB of storage. WAV is a standard format for audio recording, and the uncompressed files will retain the

detail required for any acoustic analysis in future work. The WAV files will be used for transcription, acoustic analysis and archival storage. For use in the online resources, audio recordings will be converted from WAV to MP3 and OGG Vorbis formats to enable download and also streaming via the HTML5 audio tag across a wide variety of browsers. Interviews will be transcribed using the free and open source ELAN linguistics annotator and stored in the software's XML based EAF format. Transcriptions will be synchronised to the audio files to enable real-time highlighting of transcription sections during audio playback. Transcription files will be converted to HTML for inclusion in the online resources and PDF versions will also be made available for download. Publicly available versions of the audio files and transcripts will be anonymised appropriately.

The textual data for the project (for example the 200-point questionnaires, of which there will be 488) will be stored in a relational database and the online resources will be generated as HTML5 web pages and styled using CSS3. The API developed for the project will enable the survey data to be accessed via HTTP GET requests and returned as JSON or XML data. Data will be published on the website under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license.

2b: Hardware and Software

The technical infrastructure for the project will make use of free and / or open source software where possible. Audio recordings will be made using Tascam DR-100MKII Portable LPCM Stereo Recorders and Audio-Technica AT803 Omnidirectional Lavalier Condenser Microphones. These provide the necessary quality for a legacy project of this nature. The WAV files will be converted to MP3 and OGG Vorbis using the free, open source and widely adopted Audacity editor. Transcription will be carried out with the ELAN annotator, as mentioned above. Analysis of acoustic phonetic details of the sound files will be undertaken where necessary using the free and open source PRAAT software, the standard application in the field.

The underlying data will be stored in a free and open source MySQL relational database management system. The online web resource and the CMS will be developed using the PHP server-side scripting language, most likely using an open source web development framework such as Symfony. The map interface will use the freely available Javascript-based Leaflet API, an open source alternative to the Google Maps API. Alternative map-based approaches, such as setting up a dedicated GIS server, were investigated but it was decided that Leaflet would be better suited to a project whose outputs are almost entirely web-based, providing as it does an intuitive and widely recognised user interface and a fast, feature-rich programming interface that allows layers such as choropleth maps to be easily created. Additional client-side scripting will make use of the jQuery Javascript library. The project blog will be published via a local instance of the Wordpress open source web publishing tool.

2c: Data Acquisition, Processing, Analysis and Use

An initial project website, featuring a blog and other project information will be created by the SD and launched during the first month of the project. Although technical requirements will evolve over the course of the project an initial consensus on what should be developed is important and during the first month the SD will also work in collaboration with the PI, CIs and RA to establish a preliminary requirements document. The underlying database and a first iteration of the CMS will be released in the third month of the project (March 2015). Further updates to the CMS will take place incrementally throughout the course of the project. Development of the publicly available online resources will commence towards the

end of the first year of the project (November 2015), and prototype versions will be released for feedback in January 2016. The online resources will be publicly released in October 2016 and data will continue to be added to the resources for the duration of the project. Further incremental updates to the online resources based on user feedback will be made throughout the remainder of the project.

Interviews will be recorded by fieldworkers and the RA, who will each be trained on dialect field methods and the use of the equipment in training sessions at Glasgow in May 2015. Questionnaires and interview notes will be compiled on paper during the course of the interviews. Recordings will be copied from the recording device to a fieldworker's laptop on conclusion of each interview and will be transferred to the project's network drive via a remote desktop connection as soon as an internet connection is available. Files will not be deleted from the recording device until they have been successfully transferred to the network drive. Fieldworkers will upload the interview metadata (time/place, interviewer, gender, age, location, sociolinguistic details), field reports and questionnaire answers directly to the CMS when an internet connection is available. The CMS will automatically notify the RA when a new interview has been added to facilitate project management. The RA will be able to view all records and will liaise with the fieldworkers daily during the fieldwork phase, reporting back to the PI at weekly meetings. The RA will provide the fieldworkers with feedback throughout, and they will have the opportunity to meet with the RA and PI at any point to discuss issues which need immediate attention. Informants will fill in a consent form to release copyright and agree to make the recordings publicly available under a Creative Commons (CC-BY-NC-SA) licence as mentioned above; the form will also ask them to enter their contact details in order for the project to retain contact. They will also be provided with a debrief pack which will outline the project.

Fieldworkers will transcribe their own interviews using ELAN once they have completed their fieldwork. Backup of a fieldworker's laptop to project external hard drives will also take place at this stage. The separation of recording and transcription is intended to ensure that fieldwork is completed swiftly and effectively. Once judgement interviews are transcribed, the judgement data and transcriptions will be uploaded into the CMS. Grammaticality judgement data will be scored on a 7-point Likert scale; informants will be given warm-up examples and training on scoring judgements in order to ensure that they understand the task. Each entry for a judgement of a given syntactic phenomena by an informant will be associated with the relevant metadata in the CMS, as well as any relevant comments from the interviews by the informant. The fieldworker will also upload compressed versions of the audio files to the CMS. Once a fieldworker has completed work on a record s/he will be able to sign it off in the CMS, which will notify the RA who will then cross-check the data, report to the PI regarding any issues that arise, edit or return the record to the fieldworker if required and then mark the record as published, at which point it will automatically be added to the publicly available online resources.

The online resources, the CMS, the compressed audio files and the underlying MySQL database will be located on webservers managed by University of Glasgow IT Services and running Apache. The webserver will be backed up nightly to an Ultrium LTO2 unit located remotely from the server. The project archive, which will house the uncompressed audio files, will be stored on an Active Directory network, supported by 4 domain controllers, located in two separate 'server' rooms at each end of campus. Each server has a RAID disk subsystem and is backed up nightly to devolved backup systems. Both server rooms are protected by both UPS and generators. The backup system creates and maintains two copies of each system state backup which are held on near-line disk, on-site tape and off-site tape. 7 versions of each AD state are retained for 90 days. The database schema,

system specification and procedures for data entry and management will be described in a detailed set of documents.

Section 3: Technical Support and Relevant Experience

Technical support will be provided by the SD, who will be based at the School of Critical Studies (SCS) at the University of Glasgow. The SD has over 12 years of experience developing digital humanities based websites (listed here) and has extensive experience developing websites using PHP, jQuery and other Javascript libraries, MySQL, and geospatial APIs. Support for web servers and network infrastructure will be provided by IT Services at the University of Glasgow. There are multiple developers attached to the SCS and the risk of the SD being unable to work on the project is mitigated by the existence of other developers with appropriate skills who will be able to provide the necessary technical input if required.

In considering the technical implications of this project advice has been sought from technical experts from relevant projects and organisations, such as Glasgow's IT support services, Patrick McCann, the CMS developer for Glasgow's Archives and Special Collections, Chris Fleet, the senior map curator at the National Library of Scotland, Matthew Barr and Graeme Cannon of Glasgow's Humanities Advanced Technology and Information Institute, who have previously developed the technical infrastructure for projects built around online maps and substantial audio datasets, and Joy Davidson, associate director of the Digital Curation Centre.

Section 4: Preservation, Sustainability and Use

4a: Preserving Your Data

The project data will be preserved both during and after completion in a dedicated network area on the University of Glasgow's Active Directory network, with backup procedures as described in 2c. This preservation dataset will include: an SQL 'dump' of the complete database (periodically updated); exports of the data created by the API in XML and JSON formats; uncompressed audio files; system documentation including: database/metadata schema, full description of its functionality and copies of online user instructions. The project website will also be securely maintained after completion as described earlier. Academic papers created by the project will be added to the University's 'Enlighten' repository.

4b: Ensuring Continued Access and Use of Your Digital Outputs

The University of Glasgow has a policy to maintain digital resources resulting from research projects indefinitely. A number of older humanities research projects have been re-coded or updated substantially by technical staff in the College of Arts in recent years to accommodate migration to new servers or changes to existing server technology. Continued management and maintenance of the website will become the responsibility of the technical staff within SCS after project completion.

The choice of open, non-proprietary formats such as XML and widely established archival formats such as WAV will help ensure long-term access to the project data. Features based on JavaScript will degrade gracefully and alternative paths to content will be offered, for example via the API to facilitate reuse of the data far beyond the original aims and objectives of the project. Publishing the data under a Creative Commons license will also help facilitate this. The project databases will also be made available through the searchable Edisyn network website.