

From Spoken Word to Digital Corpus: The Calum Maclean Project

John Shaw and Andrew Wiseman

University of Edinburgh

The Calum Maclean Collection (www.calum-maclean-project.celtscot.ed.ac.uk) is a searchable, standards based catalog of a collection of Scottish Gaelic oral narrative that was developed between 2005 and 2009 with major research funding from the AHRC (Arts and Humanities Research Council). It is one of a series of projects begun early in the past decade to make folklore materials in Scotland more widely available to the public through the uses of digital technology. Implicit in such initiatives has been the intention to develop multiple social applications of archived folklore materials; throughout the developments, the two primary aspects of technological applications have been a wider promotion of folklore materials through technology, and the enhancement of Scotland's main folklore archives. The major share of the various projects' activities and the funding has gone toward the institutional, social, and promotional aspects, but the size and value of the archival collections themselves have presented a strong case for the applications of technology for research purposes. Ways in which such archival collections in Scotland could be further developed became clear during a visit in 2002 to the National Folklore Collection at UCD (University College Dublin) with a view to assessing the research potential of their extensive collection by the Scottish folklorist Calum Maclean (1915-1960). The importance of the collection to ethnologists is uncontested, but until that time very little of it had been made available to researchers. Maclean's transcriptions were meticulous, written out in a clear hand, which gave rise to the possibility that the entire collection would not only be useful in a digitized form, but could be made searchable, with additional information and comments for the use of researchers in a range of disciplines. Subsequent negotiations with UCD secured their full support for a project to develop the collection and funding was awarded by AHRC late in 2004.

The bulk of the research materials at the centre of the project were collected in the West Highlands and Islands of Scotland (primarily in the Outer Isles) between 1946 and 1951 by Maclean during his employment as a full-time field collector with the IFC (Irish Folklore Commission, later to become the National Folklore Collection) in Dublin and then when he was at the School of Scottish Studies, University of Edinburgh, from 1951 until 1960. The main collection consists of Maclean's written transcriptions from wax cylinder field

recordings, which in the course of his fieldwork were shaved and re-used as an economy measure. The Collection is bound in 24 volumes (10,511 hand-written pages, approx. 2.1 million words), of which 19 contain Scottish Gaelic transcriptions, with the recording information provided on a standard form at the beginning of each recording session. The remaining volumes (c. 1850 pages) are Maclean's field diaries covering this period and written in Irish and Scottish Gaelic. They furnish a valuable context and commentary for the ethnographic materials in the main collection, as well as his own illuminating observations on ethnology and ethnologists. In addition to the materials made available by UCD, a smaller collection of transcriptions from audio-tape (10 volumes; 2224 pp., c. 440,000 words) from the School of Scottish Studies (SSS) Archives, University of Edinburgh, was also incorporated into the project. Individual reciters' names are listed by volume and page number in both collections, but no information (e.g. Aarne-Thompson/ATU international tale-types) is accessible through a catalog. The contents are primarily long folktale texts along with more than 300 song texts, and full-length autobiographies of two major Scottish Gaelic storytellers. The greater part of the items were recorded in the Outer Hebridean islands of South Uist, Benbecula and Barra, strongholds of Gaelic tradition and a primary focus for field-collectors since the century. In addition to being the most prolific collector in the history of Scottish ethnology, Maclean was a highly-trained ethnologist and a competent, careful worker. Taken together the collections (including 6 bound volumes of Irish material from Connemara) constitute the entire known corpus of Maclean's field transcriptions, amounting to around 4000 items, or 2.5 million words.

The primary *aims* of the AHRC funded project were to convert the collection of field transcriptions by Calum Maclean, as well as his diaries, into a format compatible with modern research methods to serve as an effective and flexible resource for future research within a number of disciplines; and to ensure that his vast legacy would be available to a wider audience. In this connection, the following *objectives* were identified:

- To create a digitized *literatim* electronic corpus based on the entire transcribed collection;
- To create the largest searchable digitally-based collection of Gaelic oral narrative and concurrently the largest known written electronic corpus of oral Gaelic prose;

- To provide an online, searchable catalog designed to accommodate diverse research needs, and to provide secure and managed access to the resource;
- To make the collection more widely accessible to researchers in ethnology (particularly folktale specialists), Celtic studies, linguistics, anthropology and oral history, including those without a knowledge of Gaelic;
- To encourage research collaboration, through developing the resource, with other research institutions.

Following consultation with University of Edinburgh technical staff, the development of the Collection was set out in four consecutive stages: *pre-project (incorporating a pilot project); training and orientation; digitization/markup/cataloging; dissemination/maintenance.*

The first stage of the digitization process was to scan images either from microfilm of the original field transcriptions or from the original notebooks to TIFF (Tagged Image File Format). These images were then used for the production of texts which were re-keyed (using a double-entry method to errors) into an XML template. Preliminary pilot work had confirmed that re-keying of the texts could be carried out to an accuracy of 99.95%. In the following stage, TEI-Lite (an internationally recognized standard for the production of electronic texts) markup was added to the texts in two distinct phases. During the double re-keying process a minimal but valid document instance was produced, with a TEI header incorporating generic elements and defining the structure of the body of the text from the headings and paragraphs in the handwritten source. Such files formed the raw data which acted as the foundation for more detailed analysis and work. In the subsequent phase, further markup was carried out in Edinburgh by two Gaelic-speaking researchers. The researchers were required to possess a thorough knowledge of Gaelic language and Scottish ethnology; a practical acquaintance with the School of Scottish Studies (University of Edinburgh) folktale archive and tale classifications; a familiarity with the international academic literature of the folktale; and finally the requisite level of computer skills. Their responsibility was to add markup to the TEI header to categorize the text, and to add to the body text improved interpretative and analytical markup of the material itself based upon their academic and cultural knowledge. Specifically, markup of the text involved insertion of tags for names, places, titles, contractions, foreign words, emendations, notes and so on; replacing entities

such as accents, contractions, and special symbols to accord with their ISO equivalents; assigning motifs where they occur in the main text and to the summary; assigning a genre to a given text; assigning a taxonomic classification to a text if it had been identified either as an International Tale (AT/ATU), a Migratory Legend (ML) a Witch-type (Wi) or Fairy-type (Fa) legend; assigning keywords to the text; and parsing all marked up documents against the TEI Lite-DTD (Document Type Definition) to ensure they were valid and well-formed.

Classification of items, other than those under recognized international systems, was carried out in consultation with similar or parallel digital folklore projects (see below). The most suitable genre classification for each item, generated from an in-house created list of hierarchical descriptions, and moving gradually from the general to the particular, was identified. All the metadata – such as the information concerning the informant’s details, date, place, provenance (of the material if available), etc. – were then added to the header. A proforma used by the IFC was the source for metadata creation which, in most cases, had been appended to each section of the manuscripts for each of the individual informants. All such metadata are vital for information retrieval and thus searches may be based on title, name of reciter, location, recording date, classification or keywords. Work on the project involving the two researchers was carried out entirely through Gaelic. A guiding principle of the markup procedure was to approximate as faithfully as possible the original documents in order that each notebook could be maintained along with its content. The features of XML are such that they provide flexibility in order to convey emendations such as deletions, supralinear additions, marginalia and so on which reflect, to some extent, the immediacy of the original transcriptions.

Once completed, the marked up texts were served up to an application built by University of Edinburgh technical staff. This involved loading and storing them in files on an XML capable database. A web interface was then developed which was *specifically* designed to query the texts stored in the database, and the results of any given search were then rendered for display using a CSS (Cascading Stylesheet) which published the resultant files in a web browser. As stated, the process of digitization began with the production of TIFF images and the resultant derived JPEG (Joint Photographic Experts Group) file format was used to make the fieldwork transcription images available alongside the edited texts.

The website through which the collection is accessed is bilingual and the University. It consists of a homepage, universally accessible, providing information on the resource and the project, a short biography of Calum Maclean, and useful links. The second section of the website contains the database, with options and instructions for simple or advanced searches; taxonomy guide; contents list of manuscripts; handbook; and background and technical information.¹

The materials available online comprise one of the most important and extensive folklore collections in existence for Scotland, or indeed for north-west Europe. The only transcribed Scottish folktale collection to rival it in size and quality is the work of the pioneering 19th century collector John Francis Campbell of Islay (now in MSS at the National Library of Scotland) followed by the Dewar MSS Collection (also 19th century and containing some 1.3 million words and housed at Inveraray Castle). With the demise of storytelling in the Outer Hebrides and in the West Highlands over the last half-century and more, the materials are now irreplaceable. Given its size, content and centrality to Gaelic tradition, the Calum Maclean Collection is of fundamental importance to future research and publications in the field of Scottish Ethnology. The wealth of international tales provides extensive material for comparative research; the hero-tales contribute toward a knowledge of storytelling from late medieval times; and the two full-length autobiographies transcribed from 20th century reciters of international standing provide a unique source for researching the contexts of oral narrative.

The advantage of the present online resource is to allow the application of modern research methods to an extensive body of ethnographical data. Digitization of the transcripts using the Text Encoding Initiative's TEI-Lite XML Schema provides the basis upon which word, subject, genre and contextual information searches can be constructed, and the English tale summaries provide effective access by folktale comparatists worldwide. The resulting digital corpus is suitable for rigorous analysis by computer, enabling a variety of research projects. The primary groups of users include those with an interest in the following subject-related areas: ethnology (especially from a Scottish and Irish perspective), narrative studies, Celtic studies, linguistics, anthropology and oral history. It is envisaged that the corpus will be used in a variety of ways, providing up-to-date tools for interdisciplinary research, e.g. analysis of

¹Requests by researchers for access to the database section can be directed to project staff (j.w.shaw@ed.ac.uk).

stylistics and folklore language register; oral formulae; word frequency; dialectology; discourse analysis; geographic distribution of folktales and motifs; historical/comparative studies; and lexicography. Additional benefits of the developed corpus may include aspects such as providing an up-to-date tool for interdisciplinary research by allowing access to an electronic catalog, where researchers can compare materials through an easy-to-use web interface, together with a flexible tool for searching. The flexible format of the texts provides an effective point of departure for a vastly increased output of research and related publications based on a central but hitherto largely unused collection. An electronic resource such as this can be easily replicated for any other folklore collections.

The direction taken in the future development of the collection will be significantly determined by opportunities to interact with similar or parallel collections in Scotland, and possibly further afield. Examples of online resources that contain relevant material for those studying Scottish ethnology include *Tobar an Dualchais/Kist o Riches*, which gives access to a wide variety of original fieldwork recordings drawn together from the archives of the School of Scottish Studies, BBC Scotland and the National Trust for Scotland's Canna Collection, and *Pròiseact MhicGilleMhìcheil MhicBhàtair/The Carmichael Watson Project*, which makes available the collection of the pioneering folklorist Alexander Carmichael (1832–1912). Similar projects outwith Scotland which are of an ethnological interest include *Struth nan Gàidheal/Gael Stream*, containing archival sound recordings, hosted by St Francis Xavier University at Antigonish, Nova Scotia; *Cainnt mo Mhàthar/My Mother's Tongue*; and, a recent innovative project, *An Drochaid Eadarainn/The Bridge Between Us*, which provides an interactive online social space specifically (though not exclusively) aimed at the Gaelic community of Nova Scotia. This last resource invites users to actively participate and share knowledge about ethnological materials, thereby embracing their social dimension. Taken together, the sources complement one another in serving from their various perspectives to conserve and disseminate hitherto difficult-to-access or culturally marginalized materials.

Since the Collection went online, it has provided an effective point of departure for the study of the folktale and folktale collecting in Scotland. Maclean's field diaries contain descriptions of his first reactions to communities where he carried out so much of his work, achieved against a background of his constant awareness that the recording of Gaelic folklore was a race against time. We are also able to observe at first-hand his relationships with colleagues in the Hebrides, which formed the point of departure for the politics of

access and ownership that emerged by the 1950s around the active interest taken by folklore collectors – some of them commercial – from outside of Scotland. There is also valuable information taken down from the reciters themselves regarding the transmission and spread of tales in Gaeldom. The apparently prodigious ability of Gaelic reciters to acquire a lengthy folktale orally has been a perennial area of interest for folklorists, together with the ways in which such major tales were (and are) capable of being absorbed into the repertoire from neighbouring cultures. In 1950, Calum Maclean transcribed an internal account that portrays a remarkable instance of how swiftly and adeptly individual storytellers could master these tales, and how easily in recent times they could be introduced into the Gaelic repertoire through re-oralisation. It is taken from the life story of the prolific Gaelic reciter Angus MacMillan (Aonghas Barrach) of Benbecula, held in the National Folklore Collection at UCD (NFC 1180: 301–548) and extending to 247 pages. MacMillan, who only had a smattering of English, describes going to hear a tailor in a neighbouring township, whose performance consisted of translating a story from an English book into spoken Gaelic for his audience. Calum Maclean in an aside hazards that the book was *Five Weeks in a Balloon* (first published in 1863) by Jules Verne. If he is right, what follows is an impressive feat indeed; the 1926 English translation of the work titled *Five Weeks in a Balloon: & Around the World in Eighty Days* is 374 pages in length. Angus MacMillan continues, “it took the tailor a week to read it [aloud]. When the book was completed I knew the story and returned home early, around nine o’clock ... My father inquired whether I was poorly and I replied that I was not. He knew very well that I had been listening to the story and he asked me to recite it until it came time to retire for the night. When we had taken our evening meal I started on the story. There was no one at home except for myself, my father, my mother and two sisters. As I continued with the story my sisters grew sleepy and retired. Soon after, my mother became sleepy and went to bed, but my father didn’t even wink – he had been constantly berating me every night for being so late in coming home. The sun was rising before I stopped reciting – it was a winter’s day – and I told him that would have to do for today ... I just kept on and went out to feed the cattle without going to bed at all.” Angus’s father rose early that evening and expected to hear the rest of the story, but Angus declined and went to rest. In the meantime his father was filling in the others on the story he had heard the night before. The following evening, after he had completed his chores and eaten, Angus took up where he had left off “and I continued until five the next morning. And when I was finished I said to him [my father], ‘Now I hope you won’t be so hard on me for being so late coming home. You yourself are just as late as I was two nights ago’”.

Printed Sources:

- Campbell 1890 [1860-62] John Francis Campbell. *Popular Tales of the West Highlands*. 4 vols. rev. edn. Paisley; London: Gardner.
- Maclean 1958 Calum I. Maclean. "International Folk-tales in the Archives." *Scottish Studies*, 2:113-17.
- Maclean 1975 [1959] _____. *The Highlands*. rev. edn. Inbhirnis: Club Leabhar.
- Verne 1926 Jules Verne. *Five Weeks in a Balloon: & Around the World in Eighty Days*. London: Dent
- Wiseman 2010 Andrew Wiseman. "Fear Beag a' Chridhe Mhòir: The Life and Legacy of Calum Iain Maclean (1915–1960)." In *Rannsachadh na Gàidhlig 5: Fifth Scottish Gaelic Research Conference*. Ed. by Kenneth E. Nilsen. Sydney, N.S.: Cape Breton University Press. pp. 280-98.
- Wiseman 2013 _____. "Creating a Digital Corpus of Ethnological Material." In *Scottish Life and Society: A Compendium of Scottish Ethnology Volume 1: An Introduction to Scottish Ethnology*. Ed. by Alexander Fenton and Margaret A. Mackay. Edinburgh: Birlinn. pp. 522-40.

Internet Sources:

An Drochaid Eadarainn /

The Bridge Between Us 2013 www.androchaid.ca

Arts and Humanities Research Council

2013 www.ahrc.ac.uk

Cainnt mo Mhàthar /

My Mother's Tongue 2013 www.cainntmomhathar.com

Pròiseact Chalum MhicGilleathain / Calum Maclean Project 2103	www.calum-maclean-project.celtscot.ed.ac.uk
Pròiseact MhicGilleMhìcheil MhicBhàtair / The Carmichael Watson Project 2013	www.carmichaelwatson.lib.ed.ac.uk
School of Scottish Studies Archives 2013	www.ed.ac.uk/schools-departments/literatures-languages-cultures/celtic-scottish-studies/archives
Sruth nan Gàidheal / Gael Stream 2013	gaelstream.stfx.ca
Text Encoding Initiative 2013	www.tei-c.org
Tobar an Dualchais / Kist o Riches 2013	www.tobarandualchais.co.uk