Use of legacy data in geomorphological research

Mike J. Smith\textsuperscript{a,*}, Saskia Keesstra\textsuperscript{b}, James Rose\textsuperscript{c,d}

\textsuperscript{a}School of Geography, Geology and Environment, Kingston University, KT1 2EE, UK
\textsuperscript{b}Soil Physics and Land Management Group, Wageningen University, Droevendaalsesteeg 4, 6708PB Wageningen, Netherlands
\textsuperscript{c}Department of Geography, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK
\textsuperscript{d}British Geological Survey, Keyworth, Nottingham NG12 5GG, UK

\textbf{A R T I C L E   I N F O}

Article history:
Received 1 October 2014
Revised 28 January 2015
Accepted 9 February 2015

Keywords:
Data rescue
Geomorphology
Drumlin
Striae
River
Land cover
Channel

\textbf{A B S T R A C T}

This paper considers legacy data and data rescue within the context of geomorphology. Data rescue may be necessary dependent upon the storage medium (is it physically accessible) and the data format (e.g. digital file type); where either of these is not functional, intervention will be required in order to retrieve the stored data. Within geomorphological research, there are three scenarios that may utilize legacy data: to reinvestigate phenomena, to access information about a landform/process that no longer exists, and to investigate temporal change. Here, we present three case studies with discussion that illustrate these scenarios: striae records of Ireland were used to produce a palaeoglacial reconstruction, geomorphological mapping was used to compile a map of glacial landforms, and aerial photographs were used to analyze temporal change in river channel form and catchment land cover.

\textcopyright 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Geomorphology deals with the form of the Earth’s surface and the processes that act upon and shape it. Landscape analysis within geomorphology often involves an inductive approach to the creation of knowledge. The observation of a physical system can be used as a basis for classifying sample sets of the environment and then generalizing this complexity to a standardized theory \cite{3}. This mode of data acquisition and knowledge generation dates back to at least the 1800s (e.g. \cite{4}), although it was not until the 20th century that more sophisticated techniques for data collection and modelling evolved (e.g. \cite{16}).

For the purposes of this paper, we start from the online Collins English Dictionary (http://www.collinsdictionary.com) definition of legacy as “something handed down or received from an ancestor or predecessor” or “surviving computer systems, hardware, or software”. Within the context of data rescue, legacy data refers to data that has been collected or compiled in the past. While geomorphologists may intuitively consider this to be 10s or 100s of years (or more) old, we deliberately leave the definition wide and will return to the implications of rescuing more recent data (1s of years).

One of the principle goals of rescuing legacy data is to allow reuse of that information so that it may find utility in the future – so-called downstream applications. The attendant risk is that the data is lost and so the utility denied. Data may be stored in three forms: (1) raw, (2) processed and (3) presented. In raw form, the data may be as-collected (or subsequently processed) into a standardized format. This is the most useful as it allows full access to, and use, of the data. In a presented form (e.g. journal article), the data may now be curated in order to illustrate the purpose of the original work and so becomes less useful. For example, tabulation of the latitude and longitude of the location of striae would be considered raw data, whereas display on a map would be in a presented form.

It is noteworthy that journal articles from the 1800s were often considerably longer than is current practice with the inclusion of tabulated data common \cite{30}, although this trend is now being reversed with the ability to include supplementary materials.

The storage format should also be considered when rescuing legacy data. Analogue formats such as journal articles, maps and handwritten notes are always accessible and therefore usable (although they may require physical access or physical recovery if there are no digital copies). However, an analogue to digital transfer is required – dependent upon the data type, this will entail varying losses in data fidelity (e.g. \cite{20}). For example, transcription of text would be expected to have high fidelity, whilst the scanning
of satellite images, photographs, or maps would involve at least some degradation from the original as-collected data.

As a result, for digital sources digital data is preferred over analogue data as the original values are stored. However, this still involves considerable risk concerning storage and re-use. In particular, the ability to read the format that the data is stored in and to access the physical storage medium that the data is stored on. The potential for short data lifespans is illustrated with the British Broadcasting Corporation’s Doomsday Project (http://www.bbc.co.uk/history/domesday/story). This ambitious project attempted to digitally capture the essence of life in the United Kingdom and present it as a multi-media archive. Yet within 15 years, the data was inaccessible (http://www.theguardian.com/uk/2002/mar/03/research.elearning) and required a range of rescue missions to both access the physical media and recover the data.

Within geomorphology, there are three principle reasons for using legacy data: (1) having access to past data in order to reuse it for investigating fundamental processes. For example, this may involve river discharge data from a gauging station in order to re-investigate flow regimes; (2) to access an historical record of a specific landform or process as recorded at a specific point in time. For example, this might involve topographic mapping of a landscape prior to inundation during the construction of a reservoir; and (3) in order to investigate temporal change in a phenomenon. For example, the evolution of an earth slide over decadal time scales or monitoring of land use change.

In this paper, we present three case studies that illustrate the scenarios outlined above and provide examples of data rescue that we have been involved in. The first illustrates the use of historic records of striae observations in understanding the dynamics of past glaciations, the second demonstrates formal publication of drumlin mapping from the grey literature, and the third example shows how historic aerial photography can be used to model changes in river channel form and assess the extent of succession of catchment reforestation. The paper uses these examples to develop discussion around the use of legacy data in geomorphology and, in particular, the challenges facing the discipline going forward.

2. Case studies

2.1. Case study 1: historic striae data for palaeoglacial reconstruction

Within palaeoglaciology [1], geomorphological mapping is commonly undertaken (e.g. [31]). These landforms encode information concerning the dynamics and locations of former ice masses [15]. Landforms can be either depositional (e.g. drumlins) or erosional (e.g. striae) [1], or a combination of both; their mode of formation allows the inference of the physical conditions necessary for their formation and therefore the likely processes that operated. Specifically for this research, striae are relatively shallow (mm’s), and short (<1 m), scratches or grooves on a rock surface by rock fragments embedded in the base of an overriding ice mass.

During a literature review of the former Irish Ice Sheet (IIS), Smith and Knight [32] consulted the early work of the Geological Survey of Ireland (GSI), who mapped hard rock and surficial geology. This work formed the First Series (1:\16000) of maps published in the mid to late 1800s. It was noted that some maps contained striae observations and that the accompanying memoirs often tabulated individual measurements (e.g. [14]; Fig. 1). This historic dataset potentially had value for understanding the IIS and was therefore researched in further detail.

Striae observations are not equally distributed around the island, in part due to the diligence and experience of individual field geologists. The information recorded in the memoirs varies, with at least the general location, orientation, and a description listed. For many observations, location was made with reference to the detailed Six Inch scale (1:\10,560) topographic mapping of the Ordnance Survey (OS) First Series maps (first published in 1837), which the field geologists likely used for reference. Unfortunately, the map projections of the First Series maps varied between counties in Ireland, adding an extra layer of complexity to data reuse.

Initial compilation involved transcribing the original data tabulated in the memoirs, totaling 2300 individual observations, before georeferencing of the dataset was undertaken. The simplest and most effective method of transcribing striae location involved locating the observation on an original Six Inch map sheet and identifying the same point on a modern 1:50,000 Ordnance Survey of Ireland map sheet, and recording a 12 figure grid reference in Irish National Grid coordinates. This was undertaken at the British Library (London, UK) which has a complete set of First Series OS maps; this is notable as the data rescue of the striae measurements required access to the OS maps, themselves a legacy dataset.

In addition to the memoir data, an additional 1400 measurements were transcribed from the GSI First Series maps, 600 from published/unpublished literature, and 700 from modern GSI observations. There is likely some overlap between the mapped and tabulated measurements from the GSI First Series maps, but they have been included for completeness.

The final dataset was collated in a relational database with the following fields:

1. Location: 12-figure Irish National Grid reference.
2. Source: full bibliographic reference.
3. Orientation: angle, in degrees, of striae.
4. Cross-cutting: record of relative age for striae that cross-cut one another.
5. Locational Accuracy: qualitative measure of accuracy based upon the source record and georeferencing.

Any use of secondary data should carefully consider the potential sources of error. This includes the accuracy of the original measurements (location, orientation, and cross-cutting), any transcription error in producing the original products, and then any transcription error during database compilation and subsequent georeferencing.

The final outcomes of the project included the construction of likely the largest single database of striae at over 5000 individual observations. These were then used in the production of a new map (Fig. 2) of observations [33] and subsequent use in the development of a palaeoglacial model for the evolution of the IIS during the Last Glaciation of Ireland [32].

2.2. Case study 2: drumlin mapping as a record of landscape

As noted in the first case study, glacial landforms can be used to infer processes that occurred in the past, and therefore, the likely dynamics and extent of former ice masses. Whilst striae encode small scale directional information as a result of erosional processes, suites of landforms that include deposition also persist in the landscape and may likewise be used to infer the direction and style of glaciation. They are formed at a range of scales and are more easily identified and mapped (e.g. [2]). The proliferation of remotely sensed data and its application within geomorphology [34] has led to the mapping of large numbers of landforms (>50,000) over regional and continental scales (e.g. [23,7]). Prior to the availability of satellite imagery, field mapping [17] was the predominant method for mapping landforms.

The second case study (Fig. 3) presents historical field mapping for the Glasgow region in west central Scotland and is important for two reasons. First, it presents a record of the landscape as it existed during the time of the survey. Subsequent anthropogenic or natural landscape change (e.g., urban development, landslides, etc.) have altered the surface, hence the mapping preserves this.

**Table of supposed Ice Strike.**

<table>
<thead>
<tr>
<th>County</th>
<th>Description</th>
<th>Strike A</th>
<th>Strike B</th>
<th>Strike C</th>
<th>Strike D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galway, 50/4</td>
<td>Inveraray, on west coast</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway, 51/2</td>
<td>Reservoir, on east coast</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway, 51/2</td>
<td>Comrie, on sea-coast</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway, 51/2</td>
<td>Loch Long, South, on many places near the road</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway</td>
<td>Blane, close to the E end of the bay</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway</td>
<td>Brunswick, a little S S of the jet</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway</td>
<td>Brunswick, in many places between the river and the shore</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway, 51/4</td>
<td>Benlathie, at Cowie River</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway</td>
<td>Cowie, at the river</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway, 52/3</td>
<td>Lough Inverness, west shore</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway, 52/3</td>
<td>Lough Inverness, Inver Lodge Island</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway, 52/3</td>
<td>Toraphong, creggan</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway, 52/3</td>
<td>Toraphong, at the N E houses</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway, 52/3</td>
<td>Toraphong, at Lough Inverness</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway, 52/3</td>
<td>Toraphong, at Lough Inverness</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway, 52/3</td>
<td>Knockdour, Lough Inverness</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway, 52/3</td>
<td>Knockdour, hill west of Lough Adour</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
<tr>
<td>Galway, 52/3</td>
<td>Knockdour, further west on old road</td>
<td>-</td>
<td>-</td>
<td>N 30 E</td>
<td>-</td>
</tr>
</tbody>
</table>

The strie in column D may possibly belong to the primary striation, as they agree with the striae of the dromedary, the bearing of the bay, and other features of the country; but so they extend with the fall of the ground into Berwickshire Bay, they have been considered as if cut by the ice of that branch of the Borderstone they branch of the Glasgow Bay glaciers. It is not evident what caused the N 30 E stria, but perhaps it is due to ice moving south.

The E and W striae are supposed to be in the Galway Bay glaciers when it was of large dimensions, or perhaps they may be of the same system as the stria in column B, defined by some local cause. However, it will be hereafter seen, there are other acts that seem also to belong to this east and west system.

Fig. 1. Example of tabulated striae observations from Geological Survey of Ireland memoirs [14].

Fig. 2. Extract of the map produced from compiling striae measurements for Ireland [33].
prior state. Second, it presents an example of the “grey” literature; mapping that has been undertaken and recorded, but not formally published and therefore not widely available.

The mapping was undertaken by Jim Rose between 1965 and 1968, as part of a Ph.D. at the University of Glasgow and subsequently completed whilst a lecturer at Birkbeck College, University of London. It covers 750 km$^2$ of the western Midland Valley, located between the higher ground of the Grampians to the north and the Southern Uplands to the south. Landforms recorded included streamlined hills, drumlins, eskers, moraine ridges, kames, kame terraces, kettle holes, outwash fans, till hummocks, and glacial drainage channels [1,25,27]. The field mapping was recorded on OS 1:10,560 topographic base maps marking breaks-of-slope that were subsequently inked in and the genesis of the landforms interpreted. The large-scale data was then transcribed on 1:25,000 OS topographic base maps [28] and then transferred to 34 individual A4 photographic sheets for archival processes (Fig. 3).

For the publication of this important archive [29], the A4 photographic sheets were scanned at 300 dpi and mosaicked together.

Fig. 3. Extract of glacial geomorphological field mapping produced at 1:10,000 scale reduced manually to 1:25,000 scale [35].
within a geographic information system (GIS). For the first time, this formed a single, seamless, geomorphological map of the study area. Individual landforms were then digitized in order to allow analysis of the landforms in future work. Interpretation of the landforms allowed a reconstruction of the glacial history of the region [26,29], while mapping from a subset of the area was used to compare the efficiency of different methods for mapping glacial landforms [35].

2.3. Case study 3: land use change and river channel form from historic aerial photographs

The most common use of legacy data in geomorphology are historical aerial photographs. While the first aerial photographs date to the late 1800s [19], with extensive development in the First World War, they were not widely used until the Second World War. After the war, aerial photographs were acquired for multiple purposes, in particular for the monitoring of vegetation/forest cover. This case study of the Dragonja catchment (91 km²) in southwest Slovenia, shows how historical aerial photographs from 1954, 1975, 1985, and 1994 were used to compare past land cover with a field survey completed in 2002 [12,9]; Fig. 4). The aerial photographs were scanned at 300dpi to allow digitization and subsequent interpretation. The 1954 photographs were reproduced from the original aerial imagery and were of poorer quality in comparison to the 1975, 1985, and 1995 imagery. A comparison of forest cover was made for each photograph indicating the reforestation succession that had occurred over this time, from 30% forest cover in 1954, to 75% in 2002, with intermediate states in the intermediate years (58% in 1975, 72% in 1985, and 70% in 1994). Apart from the extent of change in forest cover, the spatial distribution and succession of forest growth (from abandoned fields to young forest to mature forest) was monitored. The main change in forest cover was on steep slopes adjacent to the river.
channel which were previously cultivated as small terraces before being abandoned in the 1950s and 1960s. The abandoned fields initially reverted to grassland before returning to a fully forested state over a period of >30 years. From this land use change, an assessment of the change in erosion risk and sediment yield at the outlet was modelled [10,11].

In addition, these aerial photographs were used to assess variation in fluvial geomorphology of the channel. Due to the change in the sediment–water ratio, the river incised and narrowed over a period of 50 years. On-site fluvial geomorphological mapping allowed the reconstruction of the different stages of the adaptation of the sediment dynamics to the new land cover in the catchment. The aerial photographs were then used to validate this reconstruction (Fig. 5).

Fig. 5a shows that the river had a wide bare channel in 1954. It can be inferred that bedload was actively being transported, indicating high discharges at least once per year. Access to historical discharge and rainfall data allowed the completion of rainfall-runoff analysis which showed that there was a severe decline of water discharge with no significant change in rainfall [8]. Fig. 5b (1975) shows that by 1975, the channel had narrowed significantly, but remained bare, a result of a change in the sediment–water ratio. Due to abandonment of the agricultural fields, the ratio between the influx of sediment and water to the channel decreased. The abandoned fields rapidly became overgrown, which protected the soil from erosion and so reduced sediment influx to the river.

However, the amount of water discharged did not change significantly, creating the so-called “hungry water” effect [18], which resulted in an incision of the channel and an actively transporting river bed. This not only caused the river to incise, but the stabilized bars, which were previously actively being moved by the river during high discharge, to become overgrown so creating multiple channels in this reach of the river. By 1985 (Fig. 5c), and particularly by 1994 (Fig. 5d), the river narrowed and formed vegetated bars in the main channel, a result of further change to the equilibrium in sediment and water influx. Thirty years after the initial field abandonment, the newly grown forest resulted in low sediment flux into the river; however, water flux decreased due to the higher water consumption of the trees, and generating less runoff in all periods of the year [8]. This resulted in a reduction of both peak and base flows, allowing gravel bars to stabilize; growth of pioneer vegetation allowed further stabilization of the bars (Fig. 5d).

The use of historical aerial photos as legacy data in this example provide essential information for assessing land cover change. Moreover, these images can also provide valuable insight to the development of river morphology and complement field-based fluvial geomorphological mapping.

3. Discussion

To misquote from Orwell [21], “all data is equal but some data is more equal than others”. All data are inherently useful but, dependent upon the application, some data are more appropriate and inherently better suited to the purposes for which they are intended. However, while researchers can plan their data collection for a specific purpose, they may not know the potential future utility of their data. For this reason, legacy data have value, and therefore need preservation.

In this paper, we have outlined three areas in geomorphology for which legacy data may be used: (1) data reuse within the context for which it was collected; (2) access to historical information for a feature that no longer exists; and (3) understanding temporal change. These case studies provide examples, highlighting how legacy data can assist contemporary research, enabling invaluable insight to geomorphological processes.

The case studies also highlight how the rescue of analogue data-sets can require significant resources, but that analogue data are inherently accessible and simply require a digital-to-analogue transfer process. The same is not necessarily true for digital data-sets, as noted in the introduction. This has important implications for data storage, an issue which Research Councils UK (http://www.rcuk.ac.uk/research/datapolicy/) and the Library of Congress (http://www.digitalpreservation.gov) have begun to address. Research Councils UK require data deposition which allows the data underpinning research to be publicly accessed, enabling verification (where appropriate) and downstream reuse. This is a vital area of development as research funders realize they are vulnerable to data loss from projects they fund and thereby guilty of abusing their core organizational mission. However, this raises three significant issues which remain only partly resolved.

First, who is responsible for data? Where does ownership lie and who should maintain and curate the data? Research funders often strictly mandate the data ownership and curation policy [13], but unfunded research may become more ephemeral, an issue which some universities have begun to address in part through repositories for published research (e.g. http://eprints.kingston.ac.uk/) and data (e.g. http://data.bris.ac.uk/data/) [6]. For students or researchers on temporary contracts (e.g. Ph.D. students or post-doctoral researchers), there is an even greater risk of data loss as the individuals have greater mobility. At Wageningen University, Ph.D. students are required to make a data management plan (http://www.wageningenur.nl/en/Expertise-Services/Facilities/Library/Expertise/Support-training/training/training-display/Data-management-planning.htm).

Second, what digital data formats should be used and is there associated metadata (e.g. [24,5])? Metadata are crucial for data

Fig. 5. Channel narrowing clearly visible on aerial photographs taken from the area just upstream of the confluence of the Dragonja and its major tributary, the Rokava. From left to right, the photographs were taken in 1954, 1975, 1985, and 1994. Arrows indicate locations of channel narrowing visible on the photographs (taken from [12]).
discovery (and so reuse) via data repositories (e.g., data.gov.uk) and portals (e.g., ShareGeo: http://edina.ac.uk/projects/sharegeo; GoGeo: http://www.gogeo.ac.uk; Landscape Britain: http://landscapebritain.org.uk; [22]), as well allowing the user to understand data provenance and any restrictions or limitations with it. There may also be subject-specific standards that researchers are required to adhere to. Data formats are a more difficult topic and beyond the scope of this paper; however, any formats should be widely supported with the potential for longevity of access. For example, within academic publishing, publishers adhere to the Portable Document Format-Archive (PDF/A) standard which will ensure continued and long term access. However, for different data types the format will vary; for example, text, audio, video, and structured and unstructured databases will all be different. Within geomorphology (and other spatial disciplines), there is a need for a standardized spatial data format. Given the cost and effort expended in completing funded research, it is imperative that outputs from this work remain accessible to future generations.

4. Conclusions

In this paper we have provided a definition of, and context to, legacy data within geomorphology. More specifically we have outlined the principle uses of legacy data within the subject area, including (1) the reuse of legacy data to investigate about fundamental processes, (2) access to legacy data for an area that has subsequently changed (and cannot be investigated using contemporary data) and (3) in order to investigate temporal change. Three case studies are presented for which the authors were involved in the rescue of legacy data and subsequent use within a research project. The first of these involved the transcription, compilation, geolocation and accuracy assessment of striae data for Ireland with its subsequent use in a palaeoglacial reconstruction (use 1). The second involved the scanning and digitization of geomorphological maps of glacial landforms for the Midland Valley region of Scotland. This illustrates the re-use of unpublished material in the grey literature for an area that has undergone significant urbanization (use 2). The final case study involved the use of aerial photographs over four time epochs to evaluate temporal change for a river catchment in southwest Slovenia (use 3).

The use of legacy data forms an important element within geomorphological research – this not only includes historic analogue archives, but also more recent digital data (e.g., Landsat archive; http://landsat.usgs.gov) and the grey literature. Allowing the discovery of these rich data sources is important for undertaking research in the future. This may also involve data rescue, whereby data are made both accessible (available through an accessible medium) and usable (available in a digital format). It also requires existing researchers to appropriately curate the data they create and make it widely available. Open data sharing will lead to better research and societal outcomes.

Acknowledgements

We would to dedicate this paper to Jocelyn Riley (né Letzer) who was involved in some of the work used in this paper and the development of the glacier bedform concept. Sadly, she passed-away in August 2014.

References


[28] Rose J. Drumlin measurements: a test of the reliability of data derived from 1,250,000 scale topographic maps. Geol Mag 1975;112(361–371).


