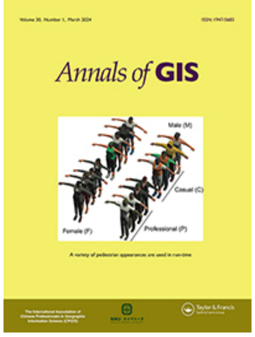


Annals of GIS

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tagi20

A century of telephony: digital capture of British telephone directories, 1880-1984

Nikki Tanu, Maurizio Gibin, Di Hu & Paul A Longley

To cite this article: Nikki Tanu, Maurizio Gibin, Di Hu & Paul A Longley (19 Mar 2024): A century of telephony: digital capture of British telephone directories, 1880-1984, Annals of GIS, DOI: 10.1080/19475683.2024.2331509

To link to this article: https://doi.org/10.1080/19475683.2024.2331509

Published online: 19 Mar 2024.

ANNALS OF GIS
https://doi.org/10.1080/19475683.2024.2331509

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

A century of telephony: digital capture of British telephone directories, 1880-1984

Nikki Tanu[a], Maurizio Gibin[a], Di Hu[b] and Paul A Longley[a,b]

[a]Department of Geography, University College London, London, UK; [b]School of Geography, Nanjing Normal University, Nanjing, China

ABSTRACT
We describe the creation of a data resource for charting the adoption of fixed line telephony in Great Britain over the period 1880–1984, of which the period 1880–1951 is available for wide research use. We first evaluate the use of various open-source software and then develop the novel teldiR open-source package to capture digitally encoded scans of all available telephone directories at ten-yearly intervals throughout this period and document the quality control checks used to establish the provenance of the resulting digital records. Our research demonstrates the feasibility of digitally encoding scans of historical telephone directories, and suggests potential analytical extensions to this work, including georeferencing of subscriber records. This research is seen not only as facilitating geographical analysis of the adoption of fixed line telephony, but also as creating a bridge between historical analysis of residential mobility, migration, and changing geodemographic structure in the years between (now digitized) historical censuses and present day digital sources. The digital resource arising from this research is available for bona fide research use.

ARTICLE HISTORY
Received 10 November 2023
Accepted 11 March 2024

KEYWORDS
Telephone directory; digitization; innovation adoption

1. Introduction

Following the inception of the public telephone service in the United Kingdom in 1879, the country's first telephone directories were introduced in 1880. Thereafter, the expanding sequence of directories reflects not just the developing complexity of the telephone network itself, but also the social and economic conditions of a changing nation. This rich source of information is in the process of being clarified and analysed at the Consumer Data Research Centre (CDRC), in collaboration with BT Archives. The CDRC is a UK Economic and Social Research Council-funded data centre, that strives to make new sources and forms of data about consumer behaviour available for research for the public good. BT Archives is part of UK telecommunications giant BT Group and has a remit to preserve and give access to BT Group's heritage, dating back to 1846: its assets include a near complete collection of original telephone directories for the UK. This paper outlines some initial findings from this collaboration, and notes some methodological issues arising from the challenging task of making a subset of the registers available in digital form for academic research.

Telephone provision began predominantly in London in 1881, and over the next decade the service coverage expanded to include four English regions in the Midlands and the North, with the exception of the North East, with services provided by separate companies in different areas. By 1889, three British telephone companies had amalgamated to form the National Telephone Company, which then went on to acquire other smaller telephone companies with regional service niches (British Telecommunications 2021). This strategy allowed the National Telephone Company to consolidate its market dominance in a market comprising a very considerable number of the most densely populated areas in England. A trunk circuit connecting London and Birmingham was opened in 1890, thus establishing routes of telecommunication between the capital city and the Midlands and some northern counties (British Telecommunications 2021). By 1901, telephone service provision had extended to cover all remaining populous settlements, including coastal settlements around England, South Wales, and Scotland.

Growth in the use of telephones in the UK during the last two decades of the nineteenth century was dependent on a complex set of interlocking factors, including technological innovation; legislative resolution and implementation; the acquisition and investment of capital; and cultural interest in and acceptance of this new medium of communication (https://www.britishtele

CONTACT Paul A Longley p.longley@ucl.ac.uk Department of Geography, University College London, Gower Street, London, WC1E 6BT, UK

phones.com/histuk.htm). Whilst the friction inherent in these transactions made for an uncertain series of outcomes, the network experienced overall growth at a level requiring frequent updates of the directories, to a point at which editions were usually produced twice a year between 1913 and 1940. BT Group Archives maintains a near complete collection of original directories for the UK from 1880, and a data sharing agreement between BT and CDRC has allowed selected issues and time periods of the directories to be digitally encoded for further analysis. The BT Group Archives collections are acknowledged by UNESCO and Arts Council England as being internationally significant and an important part of the UK's cultural and scientific heritage. While the archives contain only subscriber information for Greater London up until 1894, national and regional telephone directories emerge thereafter, containing records for subscribers across England, Wales and Scotland and later also Northern Ireland (with Northern Ireland not being included in this research). Table 1 lists the numbers of directories available for selected years between 1881 and 1951.

This paper sets out the challenges associated with the encoding of approximately 46,065 pages of historical telephone directories, scanned by ancestry.com and made available to CDRC under licence by the owner of the directories, BT Archives (Holborn Telephone Exchange, London). Scans are available for nearly every available telephone directory, but because of resource limits for the data intensive task of data encoding, we attempt only to encode directories released at decennial intervals, commencing in 1881 – a decision that allows comparison of digitized data with individual-level census data where available under special licence from the UK Data Service (Schurer and Higgs 2022). Following Information Commissioner's Office advice, the series is terminated in 1851, in recognition that living adult individuals could reasonably object to later records being made available in searchable digital form. In what follows, we detail the procedures used for extraction of text from the image scans and the reorganization of these data into tables with fields containing distinct information; we also detail the decisions and assumptions made, for reasons of transparency and to facilitate ease of data use (King 2011).

To the best of our knowledge, this is the first systematic attempt to digitally encode historical national fixed line telephony records. In the UK setting, the work plugs an important gap between the coverage of historical census records, which only become available at individual level 100 years after enumeration (thus the most recent available records are from the 1921 Census) and comparable name and address records from contemporary and current data sources. With respect to the latter, the CDRC has created and annually updated the Linked Consumer Registers (Lansley, Li, and Longley 2019; van Dijk, Lansley, and Longley 2021) every year from 1997 to the present day, through concatenation of public electoral registers and smart data sources. These historical and present-day sources each enable a rich and geographically detailed picture of population stasis and movement, since most Anglo Saxon family names have local or regional origins, and precise georeferencing enables scale-free analysis of increases in population mixing over time (Longley, van Dijk, and Lan 2021). The availability of digital national telephone subscription lists for almost the entire intervening period may enable rates of population mixing and local/regional migration and residential mobility to be calculated for the period 1931–1984. In instances in which family names are rare, it may also be able to trace entire family histories. These tasks use family names as markers of regional and local origins and require understanding of the source and operation of bias in the renting of telephone lines between different localities and family groups. This new resource for historical GIS also has potential extensions to the historical geodemographics of town and city structure – where census-based neighbourhood classifications have been built for the recent past (Gale et al. 2016; Wyszomierski et al. 2023) and for historical censuses (Lan and Longley 2019, 2021) but not for the intervening period.

Table 1. Numbers of unique directories and of subscribers available from the digital scans from 1881 to 1951.

Year	Pages	Columns	Unique Directories	Estimated Total Number of Records
1881	535	535	7	12,847
1891	1,145	1,145	3	32,471
1901	1,027	1,027	5	46,064
1911	5,612	10,938	11	717,745
1921	5,317	10,634	7	645,811
1931	12,081	28,387	11	2,199,345
1941	7,743	22,581	8	1,719,772
1951	12,605	37,815	10	2,760,380
Total	46,065	113,062	62	8,134,435

Given the extent of the data contained in the BT Group Archives, this project followed an incremental plan of work for their digitization. This process was subdivided into two phases. In the first phase, we developed proof of concept of the viability of the routines of the proposed data capture and cleaning. This work focused on telephone subscriber records from London between the years 1881 and 1901 (inclusive), where the annual volumes of records are relatively manageable for analysis. Additionally, contemporaneous digitized individual records are available for this period from national censuses of England, Wales and Scotland from the UK Data Service through the I-CeM Project (Schurer and Higgs 2022). Direct comparison of these sources allows the changing nature and coverage of telephone subscriptions in different settlements in Britain to be ascertained, potentially also broken down by social groups (Lan and Longley 2021).

Geographically, data capture in this proof-of-concept phase focused on London. Even as telephone adoption rates in major cities other than London began to climb steadily, they remained dwarfed by the numbers in London, and it is not until the late 1890s that any other city has its records in a standalone section of the telephone directories. All in all, London proved most suitable for the proof-of-concept work, given the trickle-down nature of the spread of telecommunications technology (Mahler and Rogers 1999), it was assumed that patterns of telephone uptake in London would roughly indicate how the same phenomenon would unfold elsewhere in Britain later.

In the second phase, which expanded the processing to the years after 1911, data capture also proceeded chronologically at decadal intervals. Changing directory layouts and the advent of multiple columns necessitated continual modifications to the approaches taken, and, sometimes, major changes to the processing pipeline were required to accommodate the particularities of the structures of records in a given year. To accommodate these changes, code written for data capture was developed into a series of generalized functions that could be adjusted to varying applications with only minor changes to their inputs. These functions are envisioned as a tangible output of this research, and it is hoped that other researchers looking to make sense of historical data can build upon our work.

For the early years of telephone operation in the UK, the geographical coverage of the dataset (and thus its volume) is relatively small, and it is thus feasible to trial different methods to develop processing pipelines without expending excessive computing power and time.

2. Data description and organisation

2.1. Data description

Access to the scanned telephone directories dataset was provided through the CDRC's partnership with BT Archives. This resource spans over a century, from 1880 to 1984, and includes advertisements and commercial listings in addition to data on residential subscribers published in each year apart from the missing 1893 editions. The raw dataset consists of 1.6 million scans of single pages of historical, hardcopy telephone directories that collectively occupy over 1.2 TB of storage space. Of this, the focus for digitization in this project were selected years in which decadal censuses normally took place, meaning that, proportionally, an estimated 160,000 scanned directory pages required digital encoding. Within this total, both the number of yearly issues and number of records contained per issue increased proportionally with the growth in subscriptions, consistent with increasing market penetration of landline adoption (BBC 2017). The quality of the scans varies across editions, with the print quality of images of the early editions being often poorer than in later editions.

The raw scanned images, originally delivered on a portable drive, were first transferred onto a secure server managed by CDRC and, thereafter, duplicated and stored on UCL's secure online data storage platform, DataSafeHaven (DSH). Samples of files were hosted locally on personal computers to facilitate the development of programs to process and analyse the data. Because this dataset contains digitized versions of resources which were already accessible publicly, the pre-1921 data were not considered to be sensitive, though precautions were taken to prevent access to anyone not directly named in the project or covered by the data licencing agreement with BT Archives.

As a final preparatory step before the data were processed, the scanned images were sorted based on the type of content they held. The primary objective of developing a digitization pipeline was to enable largely automated extraction of two pieces of information – the registered names of telephone subscribers and their addresses. Some pages held no relevant name and address data but some of these did provide context, such as telephone exchange maps.

Table 2 details how the raw data were provided and highlights notable variations in the types of directories, which has implications for the data processing. The scans of individual pages of each individual directory were collated into 'rolls' that were related to each other (typically by chronology) and grouped into boxes. The naming of every roll thus importantly includes a unique six-digit identifier that is always preceded by '*bt_*', as well as further information on the time of publication.

3. The data processing pipeline

A data pipeline was developed to transform residential names and addresses into a tabular format that lends itself to convenience of analysis. A flowchart of the methodology of digitization is presented in Figure 1. This pipeline encompasses a set of procedures common to all directories and its constituent steps are, broadly listed: image organization, image conversion, image preparation, Optical Character Recognition (OCR), text cleaning and separation into different fields and, finally, geocoding. While some factors that negatively impact the quality of outputs – such as the formatting of the directories and quality of scans – are non-mitigable, this research invested a great deal of effort in

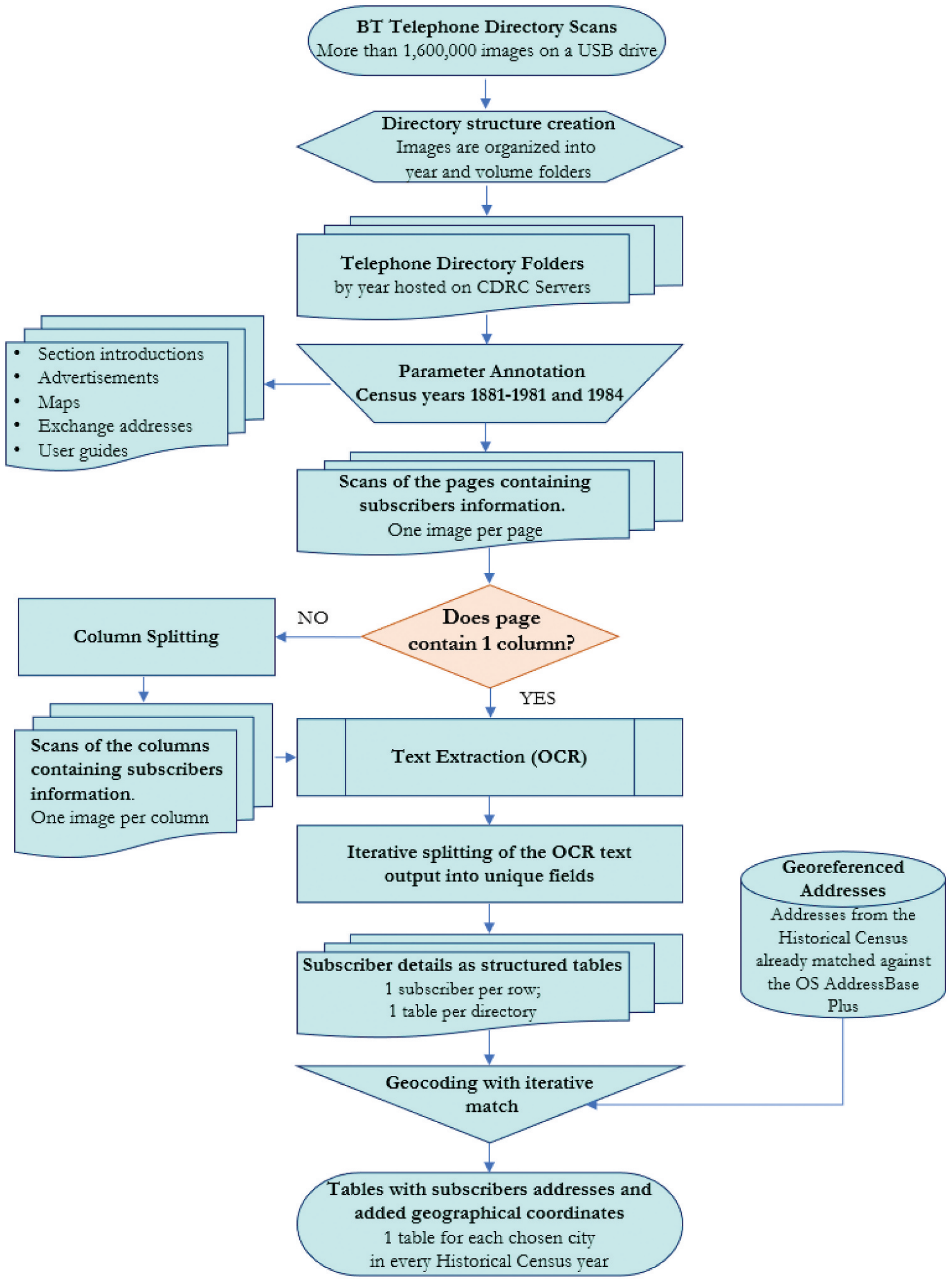

BT Telephone Directory Scans
More than 1,600,000 images on a USB drive
Directory structure creation
Images are organized into year and volume folders
Telephone Directory Folders
by year hosted on CDRC Servers
Section introductions
Advertisements
Maps
Exchange addresses
User guides
Parameter Annotation
Census years 1881-1981 and 1984
Scans of the pages containing subscribers information.
One image per page
Column Splitting
NO
Does page contain 1 column?
YES
Scans of the columns containing subscribers information.
One image per column
Text Extraction (OCR)
Iterative splitting of the OCR text output into unique fields
Georeferenced Addresses
Addresses from the Historical Census already matched against the OS AddressBase Plus
Subscriber details as structured tables
1 subscriber per row;
1 table per directory
Geocoding with iterative match
Tables with subscribers addresses and added geographical coordinates
1 table for each chosen city
in every Historical Census year

Figure 1. Flowchart of the data pipeline for processing scanned telephone directory records.

Table 2. Examples of the different types of telephone directories that were made available by BT archives for this research.

BT roll name	Type of content
bt_900008_box01_1881_jan_001	**Telephone directory** containing records of all residential and business telephone subscribers from one city (London), published in January 1881.
bt_900011_box01_1881_jul_001	**List of Professions & Trades** - only lists subscribers which are offering their professional services or as tradesmen.
bt_900591_box74_1911_jul_001 bt_900592_box74_1911_jul_002	**Full listing split across multiple directories** – each directory contains information for subscribers in Metropolitan London whose names begin with certain letters: A-H (900591) and H-Z (900592). This structure occurs with later London directories.
bt_900789_box105_1941_oct_001	**Regional directories** – beyond major cities whose numerous subscribers are enough to constitute a directory of their own, most other records are listed by region which loosely correspond to historic county delineations.

maximizing the robustness of the pipeline in the face of such inconsistencies. Hereafter, each subsection outlines a single step in the pipeline, the considerations that governed its development and the choices made between viable alternatives.

3.1. Image organisation

Prior to pipeline implementation, the pages of directories in the planned census years were manually categorized by their content to facilitate later batch processing. Table 3 details the types of pages and their respective contents.

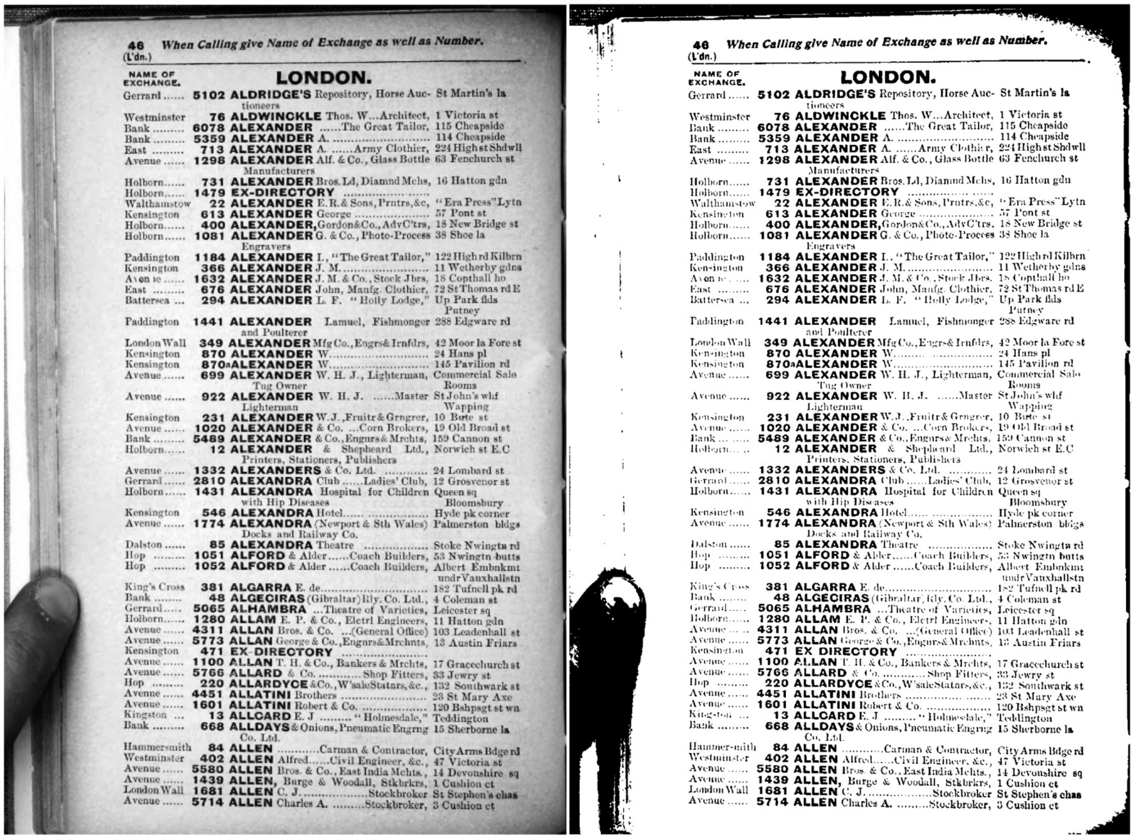
3.2. Pre-processing of images

The pipeline begins with a mass conversion of images belonging to one directory, from scans in the compressed *.j2k* format into *.tif* files, a format that is more amenable for use with common OCR software. Despite this, the scans themselves are kept as *.j2k* files for practicality: this format is an ideal compromise between minimizing file sizes and preservation of image quality – both of which are crucial for time-efficient and accurate processing of historical documents (Dueire Lins et al. 1994). All the images were subjected to cropping of borders and binarization. The former facilitates the OCR engine by defining page limits within which the engine looks for textual content. By doing this, the probability of the engine detecting irrelevant information, such as page numbers or marks on the periphery of pages, is significantly reducedWe developed a trial-and-error approach to achieve this whereby 5 random pages were sampled from each directory and the depth of white space surrounding the main text on four sides of the page was manually noted. For these pages, the minimum respective values were then used as parameters to crop and export all images from each volume. Random samples of these exports were once again checked to ensure that no essential information was cropped out of the pages. The other image pre-processing operation attempted was binarization, that is, the conversion of a set of colour pixels into either black or white pixels. In this operation, the algorithm typically determines a threshold grey value, beyond which individual pixels are changed into white pixels and below which they are turned black (Saha, Basu, and Nasipuri 2014). This threshold can either be set to apply to the entire image (global) or, alternatively, the image is broken down into smaller areas, for each of which a threshold is set (local) (Saha et al. 2014). Because nearly all the scanned pages have uniform colour tone across the image, we opted for the more straightforward global threshold. Figure 2 exemplifies how global binarisation changes the appearance of an image.

Published trials (e.g. Gupta, Jacobson, and Garcia 2007; Reul et al. 2019), indicated that heightened contrast in images would improve success in character recognition. However, during the trial on directories from the census years between 1881 and 1911, a sample of 100 binarized images that were passed through OCR software returned poorer performance than their non-binarized counterparts as measured by visual inspection of character misdetections in the outputted text files. A plausible cause may be that in historical documents that have been archived, printed text is often inconsistently shaded, either because of flaws in the original printing process or arising from wear-and-tear that naturally comes with age, thus impeding the

Table 3. Classification of directory pages prior to digital encoding of names and addresses.

Page type	Information contained
Advertisements	Full-page advertisements of products or services, containing text that is often accompanied by images
Blank	–
Introductory	Introductory information for telephone directory end-users, for instance “How to use this directory” guides
Maps	Maps (typically hand-drawn) of telephone service provision, such as the location of telephone exchanges and call offices
Subscriber Records	Alphabetically or spatially organised lists of telephone subscriber names, telephone numbers and addresses
Telephone Exchange Information	List of addresses of the main telephone exchanges from which telephony services are provided
Trades Lists	Telephone subscription information organised by profession rather than alphabetically or by subscriber location



46 *When Calling give Name of Exchange as well as Number.*
(L'dn.)

NAME OF EXCHANGE.		LONDON.	
Gerrard	5102	**ALDRIDGE'S** Repository, Horse Auctioneers	St Martin's la
Westminster	76	**ALDWINCKLE** Thos. W...Architect,	1 Victoria st
Bank	6078	**ALEXANDER** ......The Great Tailor,	115 Cheapside
Bank	5359	**ALEXANDER** A.	114 Cheapside
East	713	**ALEXANDER** A. ......Army Clothier,	224 High st Shdwll
Avenue	1298	**ALEXANDER** Alf. & Co., Glass Bottle Manufacturers	63 Fenchurch st
Holborn......	731	**ALEXANDER** Bros. Ld, Diamnd Mchs,	16 Hatton gdn
Holborn......	1479	**EX-DIRECTORY**	
Walthamstow	22	**ALEXANDER** E. R. & Sons, Prntrs, &c,	"Era Press" Lytn
Kensington	613	**ALEXANDER** George	57 Pont st
Holborn......	400	**ALEXANDER,** Gordon&Co., AdvC'trs,	18 New Bridge st
Holborn......	1081	**ALEXANDER** G. & Co., Photo-Process Engravers	38 Shoe la
Paddington	1184	**ALEXANDER** I., "The Great Tailor,"	122 High rd Kilbrn
Kensington	366	**ALEXANDER** J. M.	11 Wetherby gdns
Avenue	1632	**ALEXANDER** J. M. & Co., Stock Jbrs.	18 Copthall ho
East	676	**ALEXANDER** John, Manfg. Clothier,	72 St Thomas rd E
Battersea ...	294	**ALEXANDER** L. F. "Holly Lodge,"	Up Park flds Putney
Paddington	1441	**ALEXANDER** Lamuel, Fishmonger and Poulterer	288 Edgware rd
London Wall	349	**ALEXANDER** Mfg Co., Engrs& Irnfdrs,	42 Moor la Fore st
Kensington	870	**ALEXANDER** W.	24 Hans pl
Kensington	870a	**ALEXANDER** W.	145 Pavilion rd
Avenue	699	**ALEXANDER** W. H. J., Lighterman, Tug Owner	Commercial Sale Rooms
Avenue	922	**ALEXANDER** W. H. J. ......Master Lighterman	St John's whf Wapping
Kensington	231	**ALEXANDER** W. J., Fruitr & Grngrer,	10 Bute st
Avenue	1020	**ALEXANDER** & Co. ...Corn Brokers,	19 Old Broad st
Bank	5489	**ALEXANDER** & Co., Engnrs & Mrchts,	159 Cannon st
Holborn......	12	**ALEXANDER** & Shepheard Ltd., Printers, Stationers, Publishers	Norwich st E.C
Avenue	1332	**ALEXANDERS** & Co. Ltd.	24 Lombard st
Gerrard	2810	**ALEXANDRA** Club ......Ladies' Club,	12 Grosvenor st
Holborn......	1431	**ALEXANDRA** Hospital for Children with Hip Diseases	Queen sq Bloomsbury
Kensington	546	**ALEXANDRA** Hotel......	Hyde pk corner
Avenue	1774	**ALEXANDRA** (Newport & Sth Wales) Docks and Railway Co.	Palmerston bldgs
Dalston	85	**ALEXANDRA** Theatre	Stoke Nwingtn rd
Hop	1051	**ALFORD** & Alder......Coach Builders,	53 Nwingtn butts
Hop	1052	**ALFORD** & Alder ......Coach Builders,	Albert Embnkmt undr Vauxhall stn
King's Cross	381	**ALGARRA** E. de............................	182 Tufnell pk rd
Bank	48	**ALGECIRAS** (Gibraltar) Rly. Co. Ltd.,	4 Coleman st
Gerrard.....	5065	**ALHAMBRA** ...Theatre of Varieties,	Leicester sq
Holborn......	1280	**ALLAM** E. P. & Co., Elctrl Engineers,	11 Hatton gdn
Avenue	4311	**ALLAN** Bros. & Co. ...(General Office)	103 Leadenhall st
Avenue	5773	**ALLAN** George & Co., Engnrs&Mrchnts,	13 Austin Friars
Kensington	471	**EX-DIRECTORY**	
Avenue	1100	**ALLAN** T. H. & Co., Bankers & Mrchts,	17 Gracechurch st
Avenue	5766	**ALLARD** & Co. ....Shop Fitters,	33 Jewry st
Hop	220	**ALLARDYCE** &Co., W'saleStatnrs, &c.,	132 Southwark st
Avenue	4451	**ALLATINI** Brothers	23 St Mary Axe
Avenue	1601	**ALLATINI** Robert & Co.	120 Bshpsgt st wn
Kingston ...	13	**ALLCARD** E. J .... "Holmesdale,"	Teddington
Bank	668	**ALLDAYS** & Onions, Pneumatic Engrng Co. Ltd.	15 Sherborne la
Hammersmith	84	**ALLEN** ....Carman & Contractor,	City Arms Bdge rd
Westminster	402	**ALLEN** Alfred......Civil Engineer, &c.,	47 Victoria st
Avenue	5580	**ALLEN** Bros. & Co., East India Mchts.,	14 Devonshire sq
Avenue	1439	**ALLEN,** Burge & Woodall, Stkbrkrs,	1 Cushion ct
London Wall	1681	**ALLEN** C. J. ....Stockbroker	St Stephen's chas
Avenue	5714	**ALLEN** Charles A. ....Stockbroker,	3 Cushion ct

46 *When Calling give Name of Exchange as well as Number.*
(L'dn.)

NAME OF EXCHANGE.		LONDON.	
Gerrard	5102	**ALDRIDGE'S** Repository, Horse Auctioneers	St Martin's la
Westminster	76	**ALDWINCKLE** Thos. W...Architect,	1 Victoria st
Bank	6078	**ALEXANDER** ......The Great Tailor,	115 Cheapside
Bank	5359	**ALEXANDER** A.	114 Cheapside
East	713	**ALEXANDER** A. ......Army Clothier,	224 High st Shdwll
Avenue	1298	**ALEXANDER** Alf. & Co., Glass Bottle Manufacturers	63 Fenchurch st
Holborn......	731	**ALEXANDER** Bros. Ld, Diamnd Mchs,	16 Hatton gdn
Holborn......	1479	**EX-DIRECTORY**	
Walthamstow	22	**ALEXANDER** E. R. & Sons, Prntrs, &c,	"Era Press" Lytn
Kensington	613	**ALEXANDER** George	57 Pont st
Holborn......	400	**ALEXANDER,** Gordon&Co., AdvC'trs,	18 New Bridge st
Holborn......	1081	**ALEXANDER** G. & Co., Photo-Process Engravers	38 Shoe la
Paddington	1184	**ALEXANDER** I., "The Great Tailor,"	122 High rd Kilbrn
Kensington	366	**ALEXANDER** J. M.	11 Wetherby gdns
Avenue	1632	**ALEXANDER** J. M. & Co., Stock Jbrs.	18 Copthall ho
East	676	**ALEXANDER** John, Manfg. Clothier,	72 St Thomas rd E
Battersea ...	294	**ALEXANDER** L. F. "Holly Lodge,"	Up Park flds Putney
Paddington	1441	**ALEXANDER** Lamuel, Fishmonger and Poulterer	288 Edgware rd
London Wall	349	**ALEXANDER** Mfg Co., Engrs& Irnfdrs,	42 Moor la Fore st
Kensington	870	**ALEXANDER** W.	24 Hans pl
Kensington	870a	**ALEXANDER** W.	145 Pavilion rd
Avenue	699	**ALEXANDER** W. H. J., Lighterman, Tug Owner	Commercial Sale Rooms
Avenue	922	**ALEXANDER** W. H. J. ......Master Lighterman	St John's whf Wapping
Kensington	231	**ALEXANDER** W. J., Fruitr & Grngrer,	10 Bute st
Avenue	1020	**ALEXANDER** & Co. ...Corn Brokers,	19 Old Broad st
Bank	5489	**ALEXANDER** & Co., Engnrs & Mrchts,	159 Cannon st
Holborn......	12	**ALEXANDER** & Shepheard Ltd., Printers, Stationers, Publishers	Norwich st E.C
Avenue	1332	**ALEXANDERS** & Co. Ltd.	24 Lombard st
Gerrard	2810	**ALEXANDRA** Club ......Ladies' Club,	12 Grosvenor st
Holborn......	1431	**ALEXANDRA** Hospital for Children with Hip Diseases	Queen sq Bloomsbury
Kensington	546	**ALEXANDRA** Hotel......	Hyde pk corner
Avenue	1774	**ALEXANDRA** (Newport & Sth Wales) Docks and Railway Co.	Palmerston bldgs
Dalston	85	**ALEXANDRA** Theatre	Stoke Nwingtn rd
Hop	1051	**ALFORD** & Alder......Coach Builders,	53 Nwingtn butts
Hop	1052	**ALFORD** & Alder ......Coach Builders,	Albert Embnkmt undr Vauxhall stn
King's Cross	381	**ALGARRA** E. de............................	182 Tufnell pk rd
Bank	48	**ALGECIRAS** (Gibraltar) Rly. Co. Ltd.,	4 Coleman st
Gerrard.....	5065	**ALHAMBRA** ...Theatre of Varieties,	Leicester sq
Holborn......	1280	**ALLAM** E. P. & Co., Elctrl Engineers,	11 Hatton gdn
Avenue	4311	**ALLAN** Bros. & Co. ...(General Office)	103 Leadenhall st
Avenue	5773	**ALLAN** George & Co., Engnrs&Mrchnts,	13 Austin Friars
Kensington	471	**EX-DIRECTORY**	
Avenue	1100	**ALLAN** T. H. & Co., Bankers & Mrchts,	17 Gracechurch st
Avenue	5766	**ALLARD** & Co. ....Shop Fitters,	33 Jewry st
Hop	220	**ALLARDYCE** &Co., W'saleStatnrs, &c.,	132 Southwark st
Avenue	4451	**ALLATINI** Brothers	23 St Mary Axe
Avenue	1601	**ALLATINI** Robert & Co.	120 Bshpsgt st wn
Kingston ...	13	**ALLCARD** E. J .... "Holmesdale,"	Teddington
Bank	668	**ALLDAYS** & Onions, Pneumatic Engrng Co. Ltd.	15 Sherborne la
Hammersmith	84	**ALLEN** ....Carman & Contractor,	City Arms Bdge rd
Westminster	402	**ALLEN** Alfred......Civil Engineer, &c.,	47 Victoria st
Avenue	5580	**ALLEN** Bros. & Co., East India Mchts.,	14 Devonshire sq
Avenue	1439	**ALLEN,** Burge & Woodall, Stkbrkrs,	1 Cushion ct
London Wall	1681	**ALLEN** C. J. ....Stockbroker	St Stephen's chas
Avenue	5714	**ALLEN** Charles A. ....Stockbroker,	3 Cushion ct

Figure 2. Use of binarization (right) to increase contrast of a scanned image *(left)*. Example shows a scanned archive page from 1901.

effectiveness of running a binarized image through the Tesseract engine. The directories also commonly included pencil strikethroughs and other scribbles in grey. With binarization, the engine tended to mistakenly detect these marks as part of the printed text, thereby distorting the captured text and causing more errors. However, after extensive trials on scanned pages from different editions of the telephone directories, it was found that manual binarization of the images did not perform significantly better than the automatic binarization which could be toggled within the Tesseract OCR engine. This became the mode of processing that was eventually used.

Finally, some other image pre-processing operations were tested but ultimately did not noticeably influence the results of text recognition and were therefore omitted. First, the scans were de-skewed, or automatically rotated so that the text was vertically upright, thus aiding the OCR process (Tesseract-OCR 2021). A small but non-negligible proportion of the images from certain directories, especially from 1881, were skewed, possibly because of challenges arising from scanning individual pages bound to many other pages. However, these skews did not, in the vast majority of cases, reduce the quality of character recognition. Second, previous research (e.g. Singh and Grewal 2012) has demonstrated that dilation or erosion of text in images can help in making more recognizable characters that are either too bold or thin, respectively. This scenario relates particularly to historical documents on which ink bleeding has arisen, and to the use of now-uncommon fonts or possibly obsolete scanning methods that modern OCR engines are not adept at handling (Tesseract-OCR 2021). However, it was concluded that owing to the relative recency of the source material, both by way of the choice of typography in its production and technology used for scanning, these issues did not pose serious threats to the output. After these steps, the images were inputted into the OCR engine.

3.3. *Text capture*

Currently, a wide array of OCR software is available commercially, ranging from generic and open-source options to professional products designed for specific applications, with many proprietary OCR software programmes being variants of more all-purpose OCR engines that were tailor-made to suit particular audiences or purposes (Reul et al. 2019). However, while OCR applications on text using modern fonts and other typographic conventions (for instance, page layout) are at an advanced stage, their application on historical texts still requires considerable refinement (Reffle and Ringlstetter 2013). Following some reviews and trials of available software, the open-source OCR software Tesseract, developed by Google, was chosen because of its large user base, availability of documentation, ease of automation, range of tuning parameters available, andappropriateness of licencing arrangements.

This part of the pipeline was developed using a combination of Linux *command line interface* scripts, the Python and R programming languages and bespoke system libraries that link with one or other of these three languages. An alternative that was considered, but was ultimately discarded, was to train an OCR model specifically adapted to the historical telephone directory archives. While this would allow for more bespoke tailoring of parameters to best aid character recognition from the scans, it would also take a much longer time since the profiling of historical documents would require considerable manual input from the team of researchers (Reul et al. 2019). The final choice was thus made to use Google's Tesseract OCR engine with fine-tuning of its parameters appropriate to the particularities of each edition of the directories; this approach entailed a process of trial-and-error with almost all parameters that could be tuned in Tesseract OCR. Ultimately, it was found that some parameters had a disproportionate influence on the OCR output when tuned appropriately.

The relevance of automatic table detection pertains to the desired output of this pipeline: a data table in which every column holds a distinct field of information – crucially subscriber name, telephone exchange number and registered address – and every row represents one subscriber. With one such table derived from each page of a directory, these details were then merged to give a digitized version of each directory's records. In theory, this would be a non-issue were the OCR engine able to automatically place information into a tabular format comprising distinct fields: among other research, Gupta, Jacobson and Garcia (2007) have demonstrated the possibility of capturing text as tables from modern or even historical documents, so long as solid lines delineate different columns of information. Unfortunately, similar attempts to replicate this on the early directories up to 1911 were unsuccessful as most all directories lacked lines dividing columns of information and had either text that ran continuously, or that was inconsistently separated by punctuation such as commas and full stops. Ultimately, these difficulties made it impossible for Tesseract to detect where one field ended and another started and thus automatic table detection was unfeasible, necessitating manual intervention instead.

The second important OCR parameter to tune is the automatic detection and segregation of within-page columns, or 'page segmentation'. As aforementioned, while the quality of scans generally improves with time, an issue that becomes more prominent in later editions is that single pages become more densely packed with information packed into multiple vertical columns. Whereas in the oldest directories, every horizontal line of text referred to just one subscriber, in later years the requirement was to incorporate separation of two or more subscribers within the same row of text. Relating to this, Tesseract OCR contains 14 different modes of page segmentation including three modes that detect page segmentation with the help of algorithms and automatic detection of orientation and written script (Tesseract-OCR 2021). Results obtained when using these modes varied across pages both within and between directories, with incomplete detection of column boundaries at best and output containing a convoluted block of text at worst. We eventually found that the options for Tesseract to assume either 'a single column of text of variable sizes' or 'a single uniform block of text' worked best because they consistently generated outputs in which each record was held on a separate line, albeit without segmentation of the information from different columns. This could then be dealt with by querying specific string patterns to separate the information in each record (on a single line) into distinct fields.

3.4. The 'teldiR' data cleaning package

In order to separate fields such as subscriber names from others, such as their telephone numbers, a series of string operations were developed and executed using*'teldiR'* a library of generalized functions developed in this research in the R programming language. Not only does '*teldiR*' facilitate the division of an unstructured block of text into named fields that each contains distinct information, it also provides functionality for ease of execution of operations that are relevant before and after string manipulation. Table 4 presents illustrative functions and their applications.

'*teldiR*' thus organizes information on the capture of names and addresses by querying text patterns and then, where there is a match, truncating, splitting into two, or partially duplicating the text strings. Queries search for user-defined permutations of alphabetical, numeric and special characters present in the text capture results. Regular Expressions (RegEx) also enable the use of meta-characters to dramatically improve the flexibility with which patterns may be queried (Campesato

Table 4. Some functions of the *'teldiR'* package and examples of their use cases.

Category	Examples of use case
String operations	• split_strings and trim_string have self-explanatory uses but also allow users to restrict the querying of text patterns to only the first X characters or last X characters of a string *(where X is a number)* • collapse_RowsUpIf allows users to conditionally merge two or more rows of data, for when a record in the original directory occupied multiple lines
Process outputs of text capture	• gen_fileLST lists all text files in a given system folder for easy importation and allows users to easily select subsets of the data, relevant for instance if they are interested in records *(of a town/city)* corresponding to a defined range of file numbers
Import and export tabular records	• readin_adrDT imports already processed records (for viewing or further modification), with the option of choosing random samples from the dataset for troubleshooting or trial of new code • export_fullDT exports processed records, naming the output file with a consistent format and automatically appends the time of export for purposes of versioning
General utility	• repl_streetAbbrevs searches address fields for common abbreviations in throughfare names and replaces them with corresponding long versions • serialise_recIDs generates a unique 11-digit identifier for each record which helps when they are to be matched to other telephone records from different years, or to census records

2019). These functions allow users to swiftly and simply manipulate text strings based on a few common arguments, and to develop function-specific customization.

'teldiR' is available for public access through the CDRC GitHub (https://github.com/kinatou/teldiR). Users may find the functions in*'teldiR'* to be useful, if the directories in which they are interested are either yet to be digitized, or if they would like to improve further on the digitization of directories from specific periods or locations. In the latter scenario, the processing pipeline prioritizes generalizability of operations for application to as many different directories as possible and it is inevitable that efficacy will be lower for some directories than others. Altogether, public access to the package means that users are able to tailor-make improvements to the quality of data according to their needs, independent of this research. More widely, the package may also be useful for other unstructured or loosely structured textual data, which is not uncommon in swathes of recently digitized historical sources. The library enables rapid computation of large-data operations. Functions were written on the basis of the data table package, which has enhanced performance over basic data manipulation functions in the R programming language (Dowle and Srinivasan 2023).

4. Attribution of geographic features

One of the foremost possibilities that the dataset of digitized telephone directories affords to research is the ability to locate historical subscribers in time and space; the vast majority of records contain geographic information, specifically the address strings of residential or commercial locations and the name of the telephone exchange that serves them. Yet the process of geolocation is inherently uncertain, and attribution of geographical coordinates must be informed by the following procedures.

4.1. Preparation of address fields in the dataset

First, the subscriber information was separated into fields corresponding to different geographic identifiers, such as street name, street number or name of residence, name of area or borough, and regional district/ sub-district (where available). Values in the address field of the output data table for each telephone directory entry were further separated into subfields containing the aforementioned components through Regular Expression (RegEx) pattern querying, although not all of these elements were present in the directories for different years and separatelocations. By splitting up these elements of spatial information, the processing pipeline accrues efficiency gains and matches a greater number of shorter strings for each subscriber, rather than fewer but longer strings (Navarro 2001). This technique also enables greater geographic precision at area, street or street number scale.

The next step entails standardization of text in the address subfields (excluding street numbering) with several outcomes that ultimately harmonize the conventions in the digitized dataset with those used in the database of addresses. Firstly, many terms common to the naming of throughfares are abbreviated in the telephone directories but must be spelt in their long forms for ease of matching. For example, 'Road' is often abbreviated to 'Rd' and 'Lane' to 'Ln', while some other terms like 'buildings' have had several common abbreviations such as 'bdgs' or 'bldgs', all of which should be considered. In the case of street names, the second standardization operation converted the text strings into lowercase letters to avoid mismatches resulting simply from differing letter cases.

A third consideration is that of sets of alphanumeric characters that the OCR Engine, albeit tuned, is nonetheless inclined to confuse; this also strongly implicates the process of replacing throughfare abbreviations with their long forms as a difference of just one character can render an abbreviation undetectable by the algorithm. This relates particularly to the referencing of regional districts and (during and after World War I) regional subdistricts, the precursors to national roll-out of postcodes in 1974. In addition, much effort had to be directed at accounting common misdetections arising from similarities between some throughfare terms, the prime example being that if 'way' is mistakenly captured as 'wav', then the algorithm detects the 'av' within as short for 'avenue', resulting in an erroneous replacement.

Another field relevant to the geocoding process is the name of the telephone exchange that serves a given subscriber, but the pipeline would more than likely have placed this information in a unique field. Like the postcode for entries from London, this field allows for savings on computational power by narrowing down the number of possible matches for each subscriber. The pool of possibilities reduces from, for example, a whole city to just a number of boroughs or parishes that overlap with the telephone exchange area.

4.2. Fuzzy string matching

The test string subscriber addresses are next assigned probable geographical coordinates. To this end, fuzzy string matching procedures were adapted from Lan and Longley (2021). To achieve this, R coding was used to ensure compatibility with the rest of the processing pipeline. Each entry's street name was matched against all addresses in London in the digitized Census records of 1881, which had in turn been georeferenced by them using contemporary addresses in the Ordnance Survey's (2021) AddressBase. The AddressBase address that shared the lowest inter-string distance with each telephony directory address was then selected using an algorithm for georeferencing. When an exact or approximate street name match was found for an entry, a further search for an address with a matching street number was initiated. The resulting georeferenced addresses were accurate either to the street level or the present-day street number level. However, it is not possible to confirm whether the precision of numbered street georeferences is real or only apparent. This is because many residential neighbourhoods in the UK developed in piecemeal fashion, with new houses disrupting the house numbering system either through infill or addition of new properties at either end of the street. As a consequence, it has not been uncommon to renumber entire streets, possibly multiple times. There is also inherent ambiguity in the geocoding of addresses using historical street addresses because streets may be renumbered, renamed, or even demolished over time. Nonetheless, it is reasonable to argue that even addresses accurate only to street level will still be meaningful for geospatial analysis.

Fundamental to the georeferencing process is the choice of the string-matching algorithmwhich is used to create record linkages between the telephone directory addresses and the present-day universal OS AddressBase register. In perhaps one of its earliest incarnations, Wagner and Fischer (1974) detailed three kinds of string operations that they considered most relevant to calculate inter-string distance, that is, a measure of how similar two strings of text are. These operations were: removing a character (deletion), inserting a new character (insertion) or replacing one with another (substitution). In some variants, a fourth operation, transposition, is considered, which refers to the movement of a character either forward or backward within the string. Table 5 summarizes some common measures of inter-string distance that have been implemented in R, the coding language in which this analysis was undertaken.

Inter-string distance is calculated by summing the total number of such operations that must be applied to transform one string into the other. A distance of zero thus implies a perfect match, while a high number indicates stark dissimilarity. Further, the weights of these operations may be adjusted such that, for instance, an insertion could increase the inter-string distance by twice as much as a deletion would, should this be viewed as optimal.

Ultimately, the Hamming and Longest Common Substring methods were deemed unfeasible. The former was deemed so because frequent abbreviations of

Table 5. Key differences between common measures of inter-string distance implemented in R.

	Allowable String Operations			
String Matching Method	Deletion	Insertion	Substitution	Transposition
Hamming ** requires strings of equal length*	N/A	N/A	Yes	N/A
Levenshtein[†]	Yes	Yes	Yes	No
Optimal String Alignment (OSA) [†]	Yes	Yes	Yes	Yes, once
Full Damerau-Levenshtein[†]	Yes	Yes	Yes	Yes, multiple
Longest Common Substring (LCS)	**extracts the longest substring common to both strings, keeping the order of characters the same*			

[†] *for these methods, the weight that each kind of string operation carries in influencing inter-string distance can be adjusted manually*

Table 6. Most common detected encoding errors (London directories).

Actual Character	Common Misdetections
S	‘5’, ‘8’
E	‘B, ‘F’, ‘H’, ‘13’, ‘18’
C	‘G’, ‘O’, ‘0’
.	‘,’, ‘,’, ‘-’

throughway names and shortening of other street names (such as removing almost all vowels) would create unequal string lengths between the addresses extracted through OCR and those from the OS AddressBase. This reason, alongside the inevitable (albeit often minor) misdetection of characters, would also render it challenging to consistently detect common substrings in addresses from the two sources. The three remaining string-matching methods detailed in Table 6 were trialled on the London directories from the census years between 1881 and 1901, with allowances on string transposition being the key differentiator. The Levenshtein Distance metric, which does not allow for transpositions, was ultimately chosen; transpositions were discarded because they would be meaningful only if neighbouring characters in text strings were often wrongly ordered, a prime example being human typographic errors arising from the manual entry of a large number of addresses – which did not appear to occur in the original production of the directories. In the context of abbreviated address names and character misdetections owing to the OCR engine, the operations of deletion, insertion and substitution therefore remain more relevant.

4.3. Reconciling different administrative geographies and georeferencing

While it is theoretically possible to match each address extracted from the telephone directories to every possible address in the database, the sheer amount of computing power this would demand renders this unfeasible in practice. To circumvent this, past applications of geocoding through fuzzy string matching have used civil administrative boundaries to conduct a more targeted matching of addresses and also improve the accuracy of matching outcomes, since some popular throughway names like ‘High Street’ may have multiple occurrences, even in the same city (Lan and Longley 2019). For this reason, Lan, Van Dijk and Longley (2021), used parish boundaries when geocoding addresses from historical census data. However, while censuses record address alongside parish, addresses found in the telephone directories are not. Rather, different yearly editions have the street names and numbers of addresses accompanied by different geographical identifiers, most pertinently regional districts and sub-districts for early directories in London and telephone exchange areas for most directories as the growth of telephone service networks accelerated in the 20th century. A further necessity for geocoding the telephone subscriber addresses is therefore to devise methods to reconcile the civil administrative boundaries, according to which the AddressBase addresses are categorized, with the other geographies by which telephone directory addresses are listed. Three methods were used to geocode different directories.

The first method simply involved matching the subscribers addresses directly with AddressBase without any consideration of the historical census registration district in which it was located. This method employed fuzzy string matching to join the different parts of each address together, but was error prone because many addresses had multiple occurrences across the database.

The second method was applied only to the London directories from 1881 (when directories for no other place in Britain existed) and 1891. In these directories, addresses were listed alongside the London regions under which they fell. The region markers of each entry’s address string were used to restrict the range of possible address matches to just the AddressBase entries located in the registration districts that plausibly corresponded to the location of the extracted address. AddressBase was filtered using a spatial join between the road network and the boundaries of the historical registration districts in the relevant year in order to locate the subscriber’s address. When registration district boundaries were not available, county parish boundaries were used instead. The process of creating a correspondence list was inherently probabilistic, given that the geographies of London had changed considerably in the decades before 1881 (The Postal Museum 2021). Registration districts were chosen to conduct this matching rather than parishes because of their closer correspondence with the geography of telephone exchange service areas in Central London in 1914. In this way, the establishment of correspondence between the two geographies was limited by the granularity of the telephone exchange geographies. Because this manual effort would link every telephone exchange area to multiple civil geographical units, it made much more sense for this analysis to be done against 29 registration districts, rather than against civil parishes, which numbered 190 in 1901. Moreover, to maximize the yield of address-linked locations for ease of geocoding, many frequent misdetections of the characters occurring in London were accounted for and accordingly substituted (Table 6). Relative to the first method, this approach leveraged a threefold decrease in computer processing time and avoided counterintuitive allocations to less connected parts of cities.

This approach was adopted for the London directories in the planned census years from 1901 onwards, when telephone exchange areas become the dominant logic of geographic organization of the records of subscribers across all directories. London is chosen as the illustration here because it is by far the most complicated and presents the highest number of subscriber records to be geocoded each year. From 1901, the notation of subscriber addresses in London no longer consistently includes region identifiers, rendering the first approach unusable. A list of correspondences between the civil administrative boundaries (registration districts) and the telephone exchange geography of London was created. To achieve this, a hand-drawn map of the approximate locations of telephone exchange areas in the 1914 directories (the closest in time to 1901 that was found) was superimposed upon registration district boundaries from 1881 (see Figure 3), from which overlaps between the two maps were recorded.

Notably, owing to the density of telephone exchanges in central London, each telephone exchange was linked to a higher-than-usual number of registration districts to account for the uncertainty of not knowing where these exchange service areas actually ended. There is thus ambiguity with respect to the registration districts around, for instance, Holborn, City of London and Marylebone. Further, in comparison to the approach of individual string correction applied onto the postcodes of addresses in the 1881 directory, a second layer of fuzzy string matching was used to match the telephone exchanges to which each address belonged, consistent with registration district boundaries. Because the misdetection of characters in the telephone exchange names by the OCR engine were much more varied and irregular than those of postcode regions, employing fuzzy string matching proved to be more time-efficient in identifying exchange names which had, for instance, one or two characters misspelt or missing.

The results of this procedure for 1901 are shown in Figure 4: in London 8,167 of 10,384 residential and 11,705 of 13,276 commercial subscribers were geolocated; in Manchester 593 of 674 residential and 325 of 345 commercial subscribers were geolocated; and in Glasgow 541 of 663 residential and 165 of 169 commercial subscribers were geolocated.

The third and final approach that was applied to geocode records of subscribers from urban settlements other than London, Glasgow and Manchester was far less complicated. Thus far, we have only considered macro-level comparisons of the growth of telephone subscriptions between cities, rather than undertaking detailed mapping of the patterns that develop within cities. The approach thus entailed simply manually noting the range of pages in each year's directories that held information about subscribers in the top 50 settlements in Britain by population in 1901 (Lan and Longley

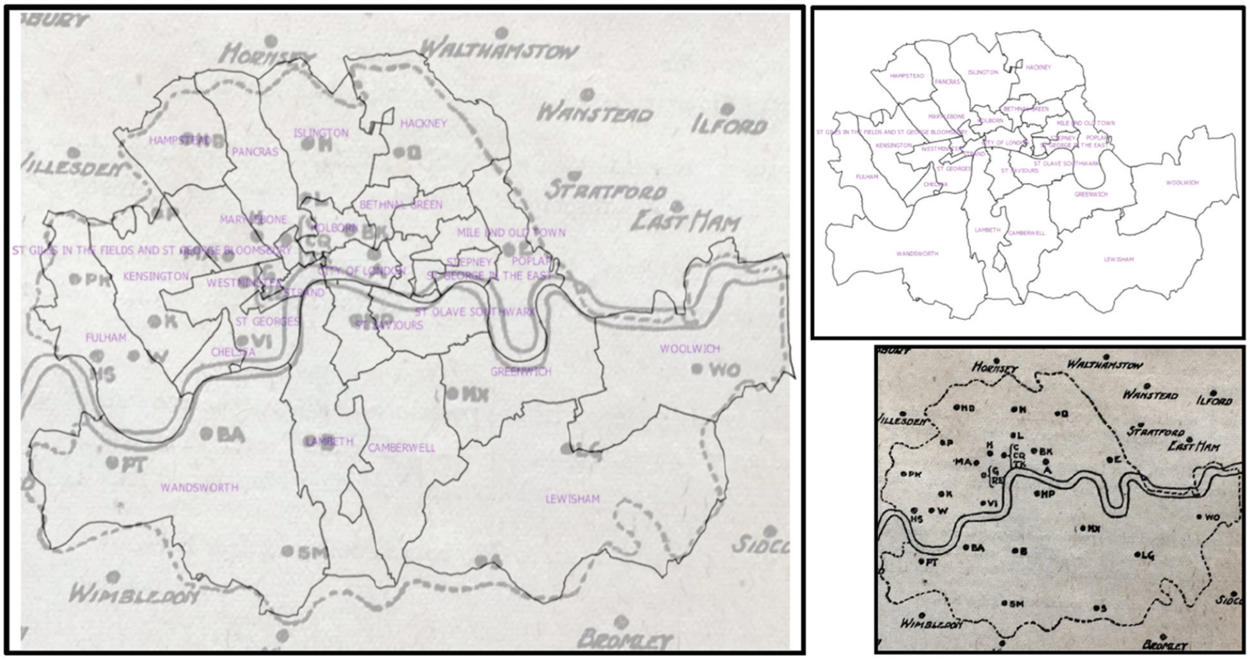


Figure 3. Composite map (left) of London's civil administrative boundaries in 1881 (right, top) and 1914 telephone exchange locations (right, bottom).

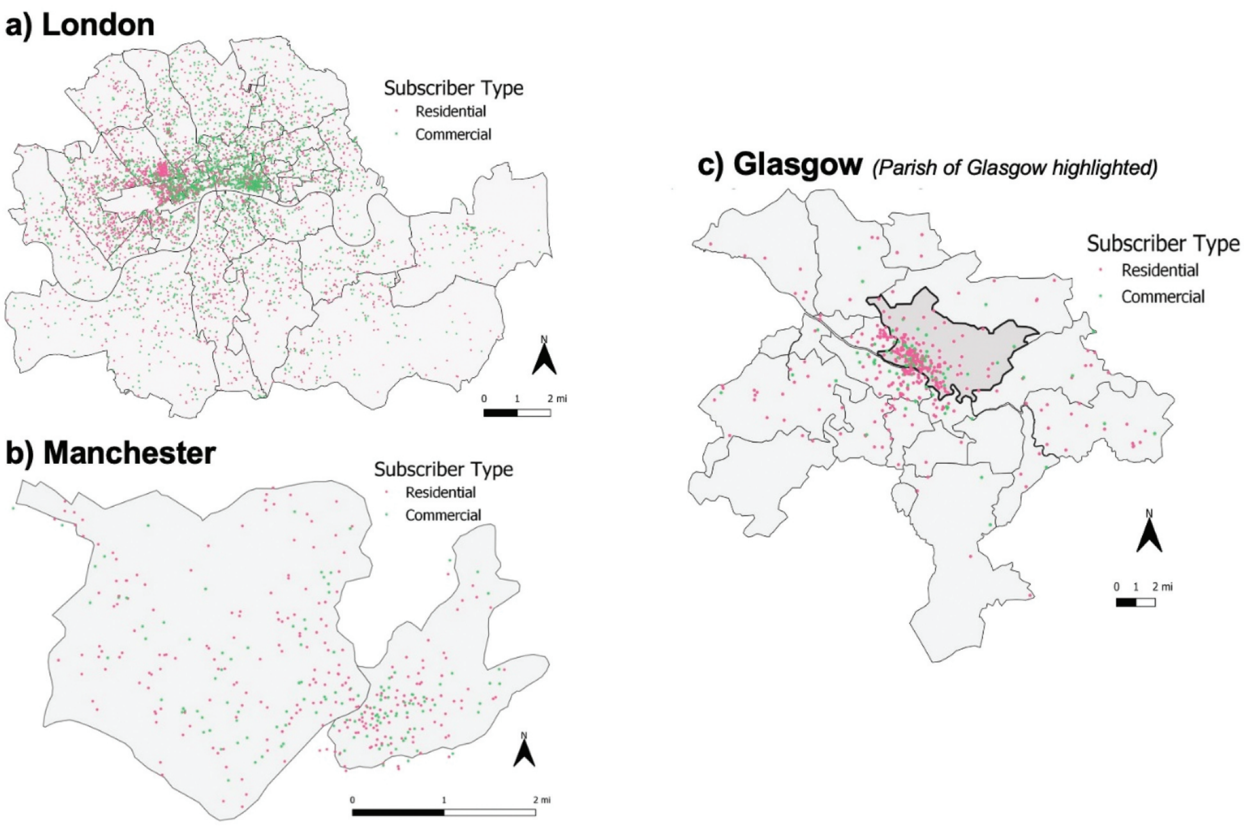


Figure 4. Locations of 1901 telephone subscribers that were geocoded in (a) London, (b) Manchester and (c) Glasgow.

2021) and then attributing to these records the settlement name.

The first two methods presented a tractable solution for the early stages of telephony adoption, since issues of street name duplication can be managed, by manual user intervention if necessary. As the network develops, however, this approach was not sustainable. In later years when multiple urban areas beyond London were added to the network, the third method was implemented, freeing the digital encoding from reliance on fuzzy matching of address strings. Instead, this third method was adapted to analyse and compare telephone adoption rates across different urban settlements by leveraging the consistent format and structure of the listings. Specifically, the number of pages associated with each freestanding settlement was used to bind the application of the method in each separate application. Average numbers of subscribers per page were calculated in order to gauge the adoption rates in each settlement. This approach was quick and intuitive, but does not provide precise geocoded information on adoption rates. Prospectively, enhancements might be made by enumerating addresses for every telephone exchange within an urban area thereby obtaining more detailed geocoded data.

5. Conclusion

This major research undertaking has demonstrated the applicability of OCR methods to archived telephone directories through a stepwise procedure, and it has demonstrated the feasibility of georeferencing the results. The reported work amounts to at least a -person year of research effort in acquiring an historical data source and developing an innovative and original data pipeline for capture of a nationally representative dataset at the level of the georeferenced identifiable human individual. This achievement is comparable and complementary to digitizing historical censuses in different parts of the developed world e.g. (Higgs and Schürer 2019). To our knowledge it is the first attempt, globally, to digitize a national series of telephone directories, and the accompanying software will be of interest to quantitative historical geographers contemplating similar endeavours elsewhere in the world. This opens up new GIScience applications in spatial analysis, modelling uncertainty and visualizing point-georeferenced population distributions.

The semi-automated procedure is robust in its application to telephone directory data with layouts that change markedly between successive updates. The pages holding key information – pertaining to telephone subscribers – are identified and systematically separated from pages with other content. The text embedded in these pages is then extracted using OCR to provide unstructured blocks of text. A series of steps is then used to restructure the text data into a tabular format, with each row containing information of one subscriber along with standardized information in subsequent fields. This processing is tailored to each unique directory layout in the decennial sample of directories. The teldiR library of modularized functions provides support for these operations. Finally, geocoding of subscriber records in select urban settlements accords them with more precise geographical coordinates, rendering them amenable to socio-spatial analysis. Uncertainty is present at each step of this pipeline, but the results provide a new and highly granular population data source, potentially for the entire first century of fixed line telephony.

While the results presented here are only exploratory, they do lay foundations for cross-disciplinary historical analysis of the individuals and businesses that are represented, the places in which they lived and the urban and regional system that evolved over a period spanning more than a century. Future work may extend data capture from the telephone directory archives. The ready-built data processing pipeline coupled with 'teldiR' enables interested parties to expand its coverage or improve on the data capture that has already been carried out. When a researcher's particular interest is focused on periods or locations for which data have yet to be digitized, they would be able to adapt our code already written to replicate the digitization process for those data; where researchers feel that the digitization already done could be improved, they could tailor the code to specific samples of records, thereby improving the quality of data capture.

This undertaking has created digitally encoded residential telephone subscriber addresses at ten-yearly intervals for the period 1881–1981, plus 1984. Consistent with UK ICO recommendations, the decennial series from 1881 to 1951 has been made available for research use through the CDRC data service. The choices that made in digital encoding are subjective yet robust, and we have documented the procedures used to devise a satisfactory outcome for most research users. Further applications of this data and methodology might include issues of migration and kinship, and the development of economic (business) and social (residential) structure of the UK's towns and cities in the late 19th and 20th centuries.

Our own research ambitions lie in developing this valuable data resource to better understand the changing geographies of residence of family groups in Great Britain at local and regional scales over the last 175 years. Telephone directory data provide a potential bridge between individual level historical census data and contemporary consumer registers (Lansley, Li, and Longley 2019; van Dijk, Lansley, and Longley 2021), which we have coupled through GIS analysis. In so doing, we also see this work as part of developing global initiatives to use link family trees (Hu et al. 2020; Jiang and Hu 2018), genealogical web services (Koylu et al. 2021) and historical censuses (Buckles et al. 2023; Helgertz et al. 2021; Schurer and Higgs 2022). This will contribute to enrichment and generalization of historical GIS data in international context.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

We acknowledge funding for this paper from the Economic and Social Research Council via the Consumer Data Research Centre grant, reference [ES/L011840/1] (Tanu, Gibin, Longley) and the Chinese 111 Program (Hu, Longley).

References

BBC Television Dial B for Britain: The Story of the Landline 2017*Timeshift*. https://www.bbc.co.uk/programmes/b08mp2l8/clips.

British Telecommunications. 2021. "'1605 to 1911 - the History of Telecommunications - Our History -About BT | BT Plc'." Accessed February, 2022. https://www.bt.com. https://www.bt.com/about/bt/our-history/history-of-telecommunications/1605-to-1911.

Buckles, K., A. Haws, J. Price, E. B. Haley, and H. E. B. Wilbert. 2023. ''Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project.'' NBER Working Paper No. 31671.

Campesato, O. 2019. *Regular Expressions: Pocket Primer*. Dulles, VA: Mercury Learning and Information.

Dowle, M., and A. Srinivasan. 2023. "'Data.Table: Extension of `data.Frame`'." Accessed September, 2023. https://r-datatable.com. https://Rdatatable.gitlab.io/data.tattps://github.com/Rdatatable/data.table.

Dueire Lins, R., M. Guimarães Neto, L. França Neto, and L. Galdino Rosa. 1994. "An Environment for Processing Images of Historical Documents." *Microprocessing and Microprogramming* 40 (10–12): 939–942. https://doi.org/10.1016/0165-6074(94)90074-4.

Gale, C. G., A. D. Singleton, C. G. Bates, and P. A. Longley. 2016. "Creating the 2011 Area Classification for Output Areas (2011 OAC)." *Journal of Spatial Information Science* 21 (12): 1–27. https://doi.org/10.5311/JOSIS.2016.12.232.

Gupta, M. R., N. P. Jacobson, and E. K. Garcia. 2007. "OCR Binarization and Image Pre-Processing for Searching Historical Documents." *Pattern Recognition* 40 (2): 389–397. https://doi.org/10.1016/j.patcog.2006.04.043.

Helgertz, J., J. Price, J. Wellington, K. J. Thompson, S. Ruggles, and C. A. Fitch. 2021. "A New Strategy for Linking U.S. Historical Censuses: A Case Study for the Ipums Multigenerational Longitudinal Panel." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 55 (1): 12–29. https://doi.org/10.1080/01615440.2021.1985027.

Higgs, E., and K. Schürer. 2019. "Data from Integrated Census Microdata (I-CeM), 1851-1911." *UK Data Service SN* 7856. https://doi.org/10.5255/UKDA-SN-7856-2.

Hu, D., X. Cheng, G. Lü, M. Wen, and W. M. Chen. 2020. "The China Family Tree Geographic Information System." In *Spatial Synthesis: Computational Social Science and Humanities*, edited by X. Ye and H. Lin, 13–37, Cham, Switzerland: Springer.

Jiang, N., and D. Hu. 2018. "GIS for History: An Overview." In *Comprehensive Geographic Information Systems*, edited by H. Bo, 101–109. Netherlands: Elsevier.

King, G. 2011. "Ensuring the Data-Rich Future of the Social Sciences." *Science* 331 (6018): 719–721. https://doi.org/10.1126/science.1197872.

Koylu, C., D. Guo, Y. Huang, A. B. Kasakoff, and J. Grieve. 2021. "Connecting Family Trees to Construct a Population-Scale and Longitudinal Geo-Social Network for the U.S." *International Journal of Geographical Information Science* 35 (12): 2380–2423. https://doi.org/10.1080/13658816.2020.1821885.

Lan, T., and P. Longley. 2019. "Geo-Referencing and Mapping 1901 Census Addresses for England and Wales." *ISPRS International Journal of Geo-Information* 8 (8): 320. https://doi.org/10.3390/ijgi8080320.

Lan, T., and P. A. Longley. 2021. "Urban Morphology and Residential Differentiation Across Great Britain, 1881–1901." *Annals of the American Association of Geographers* 111:1796–1815. https://doi.org/10.1080/24694452.2020.1859982.

Lansley, G., W. Li, and P. A. Longley. 2019. "Creating a Linked Consumer Register for Granular Demographic Analysis." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182 (4): 1587–1605. https://doi.org/10.1111/rssa.12476.

Lan, T., J. van Dijk, and P. Longley. 2021. *Urban Studies* 59 (10); 2110–2128. https://doi.org/10.1177/00420980211025721.

Longley, P. A., J. van Dijk, and T. Lan. 2021. "The Geography of Inter-Generational Social Mobility in Britain." *Nature Communications* 12 (1): 6050. https://doi.org/10.1038/s41467-021-26185-z.

Mahler, A., and E. M. Rogers. 1999. "The Diffusion of Interactive Communication Innovations and the Critical Mass: The Adoption of Telecommunications Services by German Banks." *Telecommunications Policy* 23 (10–11): 719–740. https://doi.org/10.1016/S0308-5961(99)00052-X.

Navarro, G. 2001. "A Guided Tour to Approximate String Matching." *ACM Computing Surveys* 33 (1): 31–88. https://doi.org/10.1145/375360.375365.

Ordnance Survey. 2021. *Getting Started Guide – AddressBase, AddressBase Plus and AddressBase Plus Islands. Ordnance Survey*. GB: Ronsey.

The Postal Museum. 2021. "'Postcodes'." The Postal Museum. Accessed February, 2021. https://www.postalmuseum.org/discover/collections/postcodes/.

Reffle, U., and C. Ringlstetter. 2013. "Unsupervised Profiling of OCRed Historical Documents." *Pattern Recognition* 46 (5): 1346–1357. https://doi.org/10.1016/j.patcog.2012.10.002.

Reul, C., D. Christ, A. Hartelt, N. Balbach, M. Wehner, U. Springmann, C. Wick, C. Grundig, A. Büttner, and F. Puppe. 2019. "OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings." *Applied Sciences* 9 (22): 4853. https://doi.org/10.3390/app9224853.

Saha, S., S. Basu, and M. Nasipuri. 2014. "iLPR: An Indian License Plate Recognition System." *Multimedia Tools and Applications* 74 (23): 10621–10656. https://doi.org/10.1007/s11042-014-2196-7.

Schurer, K., and E. Higgs. 2022. *I-CeMIntegrated Census Microdata (I-CeM), 1851-1911*. University of Essex: UK Data Service. https://doi.org/10.5255/UKDA-SN-7481-2.

Singh, S., and S. K. Grewal. 2012. "Text Extraction and Character Recognition Form Image Using Mathematical Morphology and OCR Technique." *International Journal of Science and Research* 3 (6): 952–955. https://www.ijsr.net/getabstract.php?paperid=2014138.

Tesseract-OCR. 2021. "'Improving the Quality of the output'." Tesseract Documentation. Accessed March, 2021. https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html.

van Dijk, J., G. Lansley, and P. A. Longley. 2021. "Using Linked Consumer Registers to Estimate Residential Moves in the United Kingdom." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 184 (4): 1452–1474. https://doi.org/10.1111/rssa.12713.

Wagner, R. A., and M. J. Fischer. 1974. "The String-To-String Correction Problem." *Journal of the ACM* 21 (1): 168–173. https://doi.org/10.1145/321796.321811.

Wyszomierski, J., P. A. Longley, A. D. Singleton, C. Gale, and O. O'Brien. 2023. "A Neighbourhood Output Area Classification from the 2021 and 2022 UK Censuses." *The Geographical Journal*. https://doi.org/10.1111/geoj.12550.