Irish Cause-of-Death Data, 1864–1921

By Ciara Breathnach, Rachel Murphy, Marco Krumrey, Daniel Sami Mitwalli, Sebastian Teumert and Tiziana Margaria

To cite this article: Breathnach, C., Murphy, R., Krumrey, M., Mitwalli, D. S., Teumert, S., & Margaria, T. (2026). Irish Cause-of-Death Data, 1864–1921. *Historical Life Course Studies, 16*, 222–235. https://doi.org/10.52024/hlcs25641

HISTORICAL LIFE COURSE STUDIES

Counting the Death.
Sources and Databases on Individual-Level Causes of Death in Historical Societies (1800–1950)

VOLUME 16, SPECIAL ISSUE 8

GUEST EDITORS
Tim Riswick
Michail Raftakis
Grażyna Liczbińska
Elena Crinela Holom

EHPS NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies was established within *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles
This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent.

Research articles
This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation, the International Institute of Social History, the European Society of Historical Demography, Radboud University Press, Lund University and HiDO Scientific Research Network Historical Demography. Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at hlcs.nl.

HISTORICAL LIFE COURSE STUDIES
VOLUME 16 (2026), 222–235, published 06-07-2026

Irish Cause-of-Death Data, 1864–1921

Ciara Breathnach — University College Cork, Ireland
Rachel Murphy — University of Limerick, Ireland
Marco Krumrey — University of Limerick, Ireland
Daniel Sami Mitwalli — University of Limerick, Ireland
Sebastian Teumert — University of Limerick, Ireland
Tiziana Margaria — University of Limerick, Ireland

ABSTRACT

Historical Irish cause-of-death data have been digitized to varying standards, and while digital copies can facilitate greater access they are not research ready. Having suffered enormous destruction over the centuries through fire and inadequate storage, Irish records are very fragmented in nature, and under-registration occurs within extant data because of poor state/subject relations and confessional differences. To reconcile absences and provide a fuller impression of historical cause-of-death, we contend that it is important to create life courses from all data types. This paper traces the development of Irish mortality data and discusses our efforts to render machine readable and interoperable data from death registration records and two decennial census datasets using innovative Low-Code/No-Code (LC/NC) solutions. Our methods and approaches will be particularly interesting to countries struggling to create Free, Accessible, Interoperable and Reusable (FAIR) data from legacy datasets that adopted a partial index linked to a PDF or Tiff of original handwritten sources.

Keywords: Civil registration, Census, Ireland

e-ISSN: 2352-6343
DOI article: https://doi.org/10.52024/hlcs25641

1 INTRODUCTION

While historical Irish cause-of-death records are digitised and accessible, they are currently geared towards genealogical purpose individual level searches, and are far from research ready. Unlike our Danish, Dutch, Norwegian and Spanish colleagues (Dahl et al., 2023; Mourits et al., 2024; Pedersen et al., 2024; Revuelta-Eugercios et al., 2021) who describe abundant transcribed sources and speak to advanced demographic methods, our target audience are scholars from countries with similar problems of fragmented records, access difficulties, legal restrictions, and varying levels of digitisation and machine readability. The aim of this article is to present the development of our open science, Low-Code/No-Code (LC/NC) solutions to expand the toolkit of transcription and handwritten text recognition methods available to humanities and social sciences' scholars. To situate our Irish case study, this section begins by outlining the development of two instruments — census and civil registration systems. We contend that the design of transcription pipelines should be sensitive to the broader socio-political contexts in which the collation of mortality data arose. The following sections 2 and 3 describe the data types together with the historiography to locate the interdisciplinary efforts of historians and computer scientists who have created innovative workflows to convert handwritten sources into machine readable data. The past and current innovations are described in sections 4 and 5, and the concluding section 6 outlines our future plans.

1.1 THE DEVELOPMENT OF CAUSE-OF-DEATH STATISTICS IN IRELAND

Ireland was under direct British rule from 1801 to 1922, under an Act of Union which made it the fourth nation in the United Kingdom (along with England, Scotland and Wales). In this timeframe two systems were established in Ireland to capture population dynamics: following a failed 1813 attempt, decennial censuses were carried out from 1821 to 1911 and civil registration of births, deaths and marriages (for all denominations) was introduced in 1864. Owing to poor state/subject relations both systems experienced implementation difficulties, thus cause-of-death statistics and their development throughout the 19th century must be viewed through a colonial lens (Breathnach, 2022a; Breathnach & Murphy, 2024; Breathnach & O'Halpin, 2012, 2014, 2021). In confessional terms Ireland was predominantly Roman Catholic, but the Protestant Church of Ireland was the official church until its disestablishment by statute in 1869. The separation of church and state began in 1871 but, for several decades after this, a high proportion of the governing elite were members of this denomination.

The early Irish census returns used an oral testimony method and are notoriously inaccurate; household forms were introduced in 1841 when William Wilde (1815–1876) first acted as a temporary assistant. Father of the internationally acclaimed writer Oscar Wilde (1854–1900), and an ophthalmologist of note, he was also a talented statistician who went beyond the remit of his temporary role to compile an additional *Tables of Deaths* section for the 1841 census report. While all censuses contained mortality data this groundbreaking report attempted to introduce a disease classification system. Drawing on William Farr's nosology, which was in use for England and Wales, local medical knowledge and vernacular interpretations of various causes, he created a total of 94 classifications (the last being "causes not specified"). More importantly it was the first to take account of the Irish language, and how scientific knowledge and ethnomedical practices combined; marasmus was known as "fairy stricken" and "back gone", which translated to *cnaoidh* or "wasting" in the Irish language (Census Report, 1843, p. vi).

Responsibility for civil registration of death, birth and marriages was placed under the auspices of the General Register Office (GRO) in 1863, which began operations in 1844 with William Donnelly as the first Registrar General (1844–1876). Initially his work was limited to the registration of non-Catholic marriages using a combination of parishes of the Church of Ireland together with the 130 districts or Poor Law Unions (PLUs) as administrative instruments. Created under the Irish Poor Law Act, 1838, PLUs later became central to local governance. On establishment each PLU erected a workhouse to administer "indoor relief". They were designed on the deterrent principles of "less eligibility" meaning that they lacked comforts, and the dietary conditions were worse than those of the poorest classes on the outside. The system was overwhelmed during the Great Famine (1845–1852) and, for our purposes, it is important to note that the links between the workhouse and excess mortality and disease were retained in collective memory until the poor law was abolished in the 1920s (Kinealy, 1997, p. 38). Despite this negative correlation, in 1851 the remit of the Poor Law was extended to incorporate a public health system under the so-called "Medical Charities Act", which divided the country into 723 dispensary districts from an expanded number of 163 PLUs. Each dispensary

employed a doctor, from whom the poor were entitled to receive free medical care under a ticketing system administered by a relieving officer (Geary, 2004). The remit of the GRO also expanded in 1851 when it became responsible for carrying out the census with William Donnelly as chief commissioner, and William Wilde now employed as an assistant commissioner (Crawford, 2003).

Given the association with Famine and general poverty, Donnelly proposed that the GRO should not employ PLU staff to conduct the work of birth and death registration, but with few alternatives in 1863 the placement of civil registration under Irish poor law was inevitable (First Report RG, 1850). Dispensary doctors became registrars and were responsible for the maintenance of birth and death registration, and the associated task of compulsory smallpox vaccination, which was also introduced in 1864. Compliance with birth registration and vaccination legislation relied on community engagement and in remote areas dispensary doctors had to deal with occasional smallpox outbreaks and hundreds of vaccination defaulters at any given time (Brunton, 2008). It naturally follows that the early civil registration records are deficient. Varying levels of life event under-reporting in all age cohorts continued well into the 20th century, where religious rites of baptism and burial traditions took precedence over civil registration of birth and death (Breathnach & Murphy, 2024; Pozzi & Kennedy, 2020a; Pozzi et al., 2020b).

Ireland was the last of the so-called "four nations" project to establish a civil registration system. Its introduction in 1864 followed that of England and Wales (1837), and Scotland (1854) and it was beset by many operational difficulties. In accordance with the 1863 Act, deaths had to be reported to the registrar by somebody present at the event or an occupier of the same dwelling within seven days; to compel greater engagement, this was adjusted to five in 1881. William Farr's 1844 nosology which delineated 145 causes of death was reprinted for Irish purposes (Statistical Nosology, 1864). Over the first five decades of the Irish GRO, the Annual Reports of the Registrar General (ARRGs) detail the considerable difficulties involved in encouraging doctors to adhere to the nosology and the updated terminology as it emerged (Breathnach & Gurrin, 2017).

As explored elsewhere the majority of the Irish population were poor and "reticent" British subjects for whom encounters with statecraft were fraught with fear and apprehension. Evidence of this weak social contract became clear on the introduction of the old age pension, when the early censuses were reclassified to provide applicants with written proof of their age. It transpired that many rural people had lived their entire lives without any interaction with the state (Breathnach, 2025; Breathnach & Murphy, 2024). The use of the Irish Constabulary (later the Royal or RIC) and the Dublin Metropolitan Police (DMP) (established in 1836) as census enumerators from 1841 may have exacerbated matters, especially when the tensions between the Crown forces and the populace are considered (Breathnach & O'Halpin, 2021; Crawford, 2003). For example, during the War of Independence (1919–1921) policemen could not undertake the scheduled census enumeration in 1921. Despite the disturbed state of the country and the imposition of martial law in Munster and parts of Leinster, the GRO was one of the few offices to continue to submit annual reports to Westminster. Consequent to partition, which was finalised in 1922, the last all-Island decennial census was conducted in 1911 and the Ministry for Finance was tasked with the duty for Northern Ireland. Detailed aggregate reports were compiled from each census exercise but only two of the household census returns survived in their entirety: 1901 and 1911. In this article we deal with the period 1864 to 1921, which includes all-Island figures for census and civil registration records.

Our work takes account of the power imbalances between the populace and the various systems of biopower and governmentality (Foucault, 2003) put in place by British administration and the fragmented nature of Irish historical life course records.[1] With a particular focus on cause-of-death, the overarching aim of our ongoing collaborative project is to create interoperabilities between the GRO data and the only fully extant census returns from pre-independence Ireland (1901 and 1911) to reconcile absences and create life course data.

1 Foucault used these terms during in his lectures at the Collégè de France, to describe the systems, institutions and instruments of power governments implemented to subjugate and control the population. Examples of such measures include the act of counting the living through census taking, and the dead, through civil registration instruments.

2 PAST AND CURRENT STATE OF THE ART IN THE FIELD

Irish demographic history lags significantly behind other European countries because of past, and indeed current, restrictions on access to records. Timothy Guinnane's *The Vanishing Irish* (1997) built on a handful of articles that caused debate and dissent between Fitzpatrick et al. (1983) and Gibbon and Curtin (1978) on the matter of Irish household formations. This body of research used samples extracted from household census forms but was not followed by further critical inquiry on Irish family structure as, in a retrograde piece of legislation (from a historian's perspective), the 1993 Statistics Act placed a 100 year embargo on access. With its implementation the majority of the extant records were rendered inaccessible for about 15 years, which had a profound impact on the development of the field of Irish historical demography.

Since 2007 both the 1901 and 1911 censuses, and what survives of those that preceded them, have been made available freely online by the National Archives of Ireland (NAI), which facilitates access to household and individual level data albeit in digital formats that cannot facilitate large scale research. With the exception of a study on Protestant and Roman Catholic intermarriage (Fernihough et al., 2015) and Connor (2021) on naming conventions, religion and fertility, scholarship on the household census returns has been constrained because the online system is limited to individual searches. However, the availability of this census data has facilitated further works by Scalone et al. (2023) on religion and child death, and Pozzi et al. (2020b) and Pozzi and Kennedy (2020a) on child and maternal mortality. We discuss some of the issues relating to linking death data with the 1901 and 1911 Irish census data in Doherty et al. (2022) and more recently in Breathnach et al. (2021, 2023, 2025). The GRO data was not available to researchers until Breathnach and O'Halpin's work on so-called "unknown and unnamed (infant) dead" (2012, 2014). Both studies were conducted under a limited data access agreement. This was followed by two funded research projects: the first (2013–2017) focused on maternal and infant mortality (Breathnach & Gurrin, 2017, 2018) but it too was limited by data-sharing constraints. Both research projects found many idiosyncrasies within the data, for example, how poor women who had concealed pregnancies risked discovery by placing infant bodies in public places to ensure burial rites. A second major discovery accrued by matching women's deaths in reproductive years with birth registration records which found that maternal mortality was underreported. These findings motivated the third project, Death and Burial Data Ireland 1864–1922 (DBDIrl, Irish Research Council Laureate Award 2017/32) which ran from 2018 to 2023 with an expanded data agreement and has yielded greater research returns that we aim to advance towards creating FAIR-compliant data using the LC/NC WAVE platform.

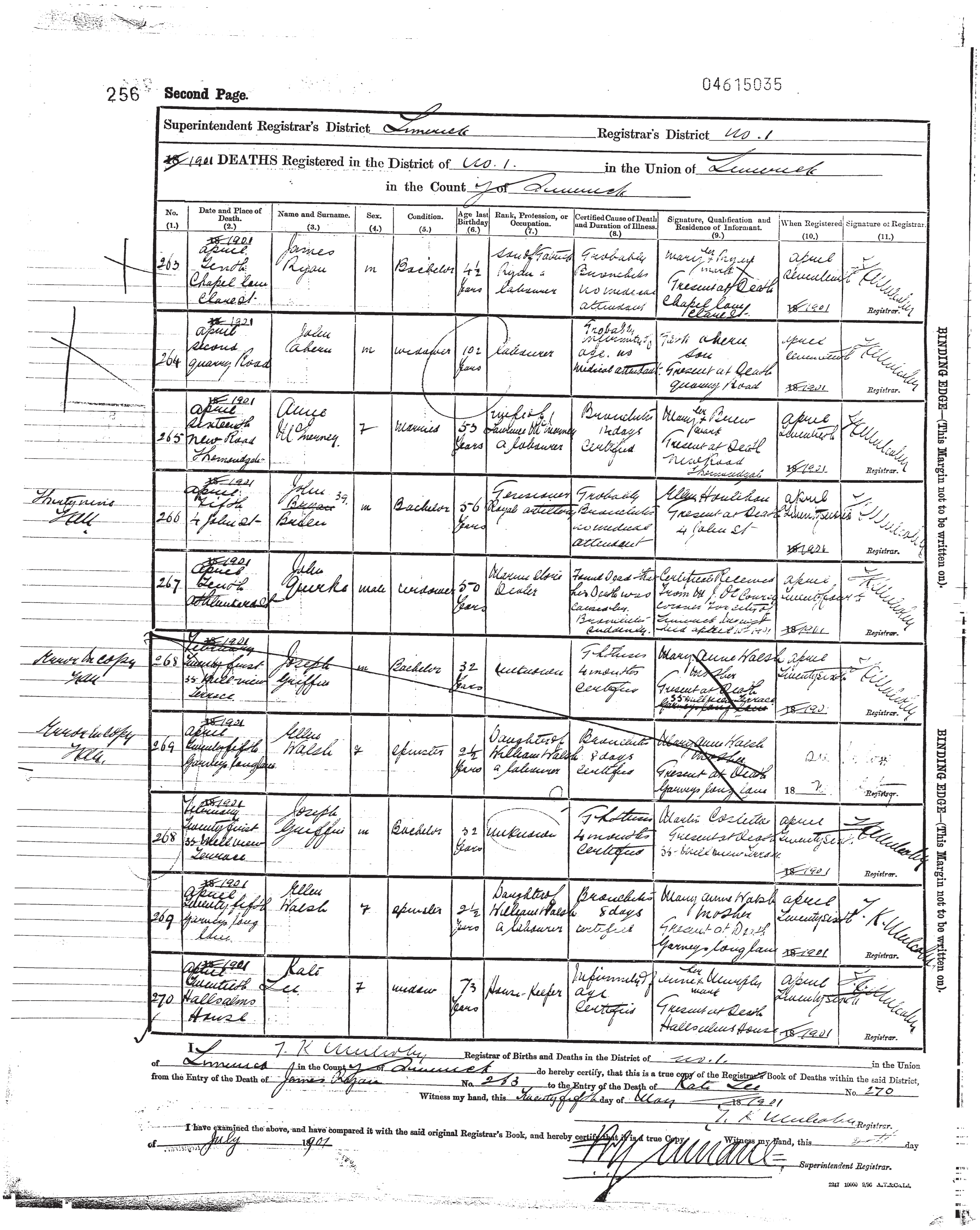
3 RECORD ACCESS AND DIGITISATION CHALLENGES

While some countries have a system of continuous registration allowing historical demographers to trace a complete life course, in Ireland (as in England and Wales, and Scotland) the life course has to be traced by individually linking GRO records with census records over 100 years old. In contrast with the publicly-available English, Scottish and Welsh censuses, which are fully extant from 1841 to 1921, Irish records are more fragmented. Apart from 1901 and 1911, which are fully transcribed and available online on nationalarchives.ie, very few of the manuscript household returns survived the ravages of fire in the Public Records Office of Ireland (PROI) in 1922, which according to the Deputy Keeper of Public Records' annual reports contained the 1813, 1821, 1831, 1841 and 1851 original census forms. Despite the lack of provision for scheduled destruction under the Public Records Act 1867, the more recent census records (1861–1891) were sacrificed in space and cost saving exercises. The twenty-eighth report of the Deputy Keeper of the PROI noted in 1896 that the 1861 and 1871 returns were destroyed (The twenty-eighth report of the deputy keeper, 1896), and anecdotal evidence suggests that 1881 and 1891 were pulped during the paper shortages of the Great War. The extant household returns and ancillary records were digitised and gradually published online by the NAI from 2007. An updated site was launched in 2024 and on 18 April 2026 the 1926 data for the 26 counties of Ireland were placed online.

Civil registration of death registers have a near complete survival from 1864 to 1970, but until 2016 the search and retrieval of General Register Office (GRO) data was a source of revenue generation: index searches on site cost €2 and a copy of each death certificate cost €4. Extensive research on these records was therefore prohibitively expensive until recently (Breathnach & O'Halpin, 2012, 2014).

Although civil registration data is now freely available online, the GRO interface relies on specificity and is limited to individual searches. In Northern Ireland the individual costs remain in place, making an all-Island study beyond 1921 less feasible (a devolved government was established by statute in 1920, and opened in May 1921). The digitisation of GRO death registration data is ongoing, with records from 1871 to 1975 freely available to search and view individually on irishgenealogy.ie. Their partial transcription for indexing purposes did not extend to cause-of-death information. Similar to several other jurisdictions, the purpose of the digitisation exercise was to facilitate individual record retrieval for official and legal purposes, and not large-scale research projects on historical cause-of-death. Our work, in cooperation with the GRO, aims to scale up full transcription to create a robust national dataset for the 32-county island of Ireland up to 1921 and for the 26 counties of Ireland until 1975 (death records older than 50 years are accessible).

Figure 1 *A page from the 1901 civil register of deaths for Limerick, District No. 1*


256 Second Page. 04615035

Superintendent Registrar's District Limerick Registrar's District No. 1

1901 DEATHS Registered in the District of No. 1. in the Union of Limerick in the County of Limerick

No. (1.)	Date and Place of Death. (2.)	Name and Surname. (3.)	Sex. (4.)	Condition. (5.)	Age last Birthday (6.)	Rank, Profession, or Occupation. (7.)	Certified Cause of Death and Duration of Illness. (8.)	Signature, Qualification and Residence of Informant. (9.)	When Registered (10.)	Signature of Registrar. (11.)
263	1901 April Tenth Chapel Lane Clare St.	James Ryan	m	Bachelor	4½ Years	Son of Patrick Ryan a Labourer	Probably Bronchitis No medical attendant	Mary Ryan Present at Death Chapel Lane Clare St.	April Seventeenth 1901	T. K. Mullally Registrar.
264	1901 April Second Quarry Road	John Ahern	m	Widower	102 Years	Labourer	Probably Infirmity of age. No medical attendant	Patk Ahern Son Present at Death Quarry Road	April Seventeenth 1901	T. K. Mullally Registrar.
265	1901 April Sixteenth New Road Thomondgate	Anne McInerney	F	Married	53 Years	Wife of Laurence McInerney a Labourer	Bronchitis 14 days Certified	Mary Breen Present at Death New Road Thomondgate	April Seventeenth 1901	T. K. Mullally Registrar.
266	1901 April Fifth 4 John St.	John Boyden	m	Bachelor	56 Years	Pensioner Royal Artillery	Probably Bronchitis No medical attendant	Ellen Houlihan Present at Death 4 John St.	April Twentieth 1901	T. K. Mullally Registrar.
267	1901 April Tenth Athlunkard St.	John Quirke	male	Widower	50 Years	Marine Store Dealer	Found dead. Her Death was caused by Bronchitis suddenly	Certificate Received from J. O'Connor Coroner for City of Limerick Inquest held 1901	April Twentieth 1901	T. K. Mullally Registrar.
268	1901 February Twenty first 35 Hill View Terrace	Joseph Griffin	m	Bachelor	32 Years	Unknown	Phthisis 4 months Certified	Mary Anne Walsh Present at Death 35 Hill View Terrace	April Twentysixth 1901	T. K. Mullally Registrar.
269	1901 April Twenty fifth Gurneys Long Lane	Ellen Walsh	F	Spinster	2½ Years	Daughter of William Walsh a Labourer	Bronchitis 8 days Certified	Mary Anne Walsh Present at Death Gurneys Long Lane		Registrar.
268	1901 February Twenty first 35 Hill View Terrace	Joseph Griffin	m	Bachelor	32 Years	Unknown	Phthisis 4 months Certified	Martin Costello Present at Death 35 Hill View Terrace	April Twentysixth 1901	T. K. Mullally Registrar.
269	1901 April Twenty fifth Gurneys Long Lane	Ellen Walsh	F	Spinster	2½ Years	Daughter of William Walsh a Labourer	Bronchitis 8 days Certified	Mary Anne Walsh Mother Present at Death Gurneys Long Lane	April Twentysixth 1901	T. K. Mullally Registrar.
270	1901 April Twentieth Hallsdalms House	Kate Lee	F	Widow	73 Years	House Keeper	Infirmity of age Certified	Annie Murphy Present at Death Hallsdalms House	April Twentysixth 1901	T. K. Mullally Registrar.

I, T. K. Mullally, Registrar of Births and Deaths in the District of No. 1 in the Union of Limerick in the County of Limerick do hereby certify, that this is a true copy of the Registrar's Book of Deaths within the said District, from the Entry of the Death of James Ryan No. 263 to the Entry of the Death of Kate Lee No. 270

Witness my hand, this Twentyfifth day of May 1901 T. K. Mullally Registrar.

I have examined the above, and have compared it with the said original Registrar's Book, and hereby certify that it is a true Copy. Witness my Hand, this day of July 1901

Superintendent Registrar.

BINDING EDGE—(This Margin not to be written on).

Source: https://www.irishgenealogy.ie/files/civil/deaths_returns/deaths_1901/05723/4615035.pdf

Figure 1 illustrates some of the issues associated with the GRO registers, which present a challenge when digitising this information in automated or manual processes. The first entry shows James Ryan (No. 263) a four-year-old child who died at home without medical attendance in 1901. His grandmother Mary's account of his symptoms was translated to "probably bronchitis" by the registrar as cause-of-death. Problems associated with the use of unqualified informants and uncertified cause-of-death are inherent in these data, and reasons for inaccurate reporting range from sensitivity to stigmatisation and poverty to poor medical knowledge, all of which complicate the way in which databases can be constructed (Breathnach, 2022a), or indeed how ICD10h can be applied to datasets (Reid et al., 2024). Two causes of death could be recorded in the registers, along with the duration of the respective illnesses. Similar to other studies, we have found that the coupling of common causes like cardiac and lung disease and the use of second cause of death and different disease durations also presents challenges (Reid et al., 2015; Revuelta-Eugercios et al., 2021). Cases of sudden or suspicious deaths (like that of John Quirke who died suddenly of bronchitis, see entry 5, Figure 1, No. 267) were to be reported to the local police station, which compiled a report for the local coroner who decided if an inquisition was necessary. If they proceeded to full inquisition, the coroner's court provided a medical verdict and registered the death (entry 5, Figure 1), and there was often a difference between cause of death in both records, especially in cases of violent deaths (Breathnach, 2022b).

4 DATABASE CREATION

That access to historical micro-data has been restricted until relatively recently means that scholars of Ireland have a great deal of catch-up to do with respect to database creation. This section describes the iterative development of our transcription infrastructure, from manual methods to an automated AI-supported pipeline. The goal throughout has been to convert handwritten death registration records into machine-readable, research-ready datasets at scale.

4.1 MANUAL TRANSCRIPTION

The construction and development of pilot databases was funded by Research Ireland as part of the DBDIrl project. We secured a copy of the index (partial) and Tiffs of the original files from the GRO and set about finding ways of converting the information into machine readable datasets. The conditions of our data agreement in 2018 did not extend to using third party operators like Transkribus, whose functionality in reading tabular handwritten registers has improved considerably in the interim (Dahl et al., 2023). Our training datasets were created through transcribathons (three-hour group sessions where individuals work together to decipher handwriting and conduct data entry). These were initially conducted with domain experts: team members and postgraduate students. The first transcribathons formed part of undergraduate and postgraduate studies at the University of Limerick, and we initially used Excel spreadsheets to enter the transcribed data. We learned a great deal about difficulty thresholds, which subsequently informed the design of a dedicated web application, the Historian DIME App (HDA) (Khan et al., 2021; Schieweck et al., 2022). In summary, HDA divided the original 11 data fields (see Figure 1) into 63 alpha-numerical components, focusing on the three different agents involved in a death registration: the deceased, the informant (a relative or the coroner), and the registrar. The software was developed using the eXtreme Model Driven Development approach (XMDD) as described in Margaria and Steffen (2020). HDA is now a legacy application due to DIME having since been deprecated, but it permitted the pre-population with the partial index received from the GRO, while the rest of the data was entered by our transcribers. Our resulting clean datasets and the efforts to create them brought the problem of transcribing millions of records into sharper focus, and the necessity to employ automated systems incorporating the use of software became more apparent.

HDA was equipped with predefined correctness checks (expressed by syntactic and semantic constraints) that ensured both the accuracy and plausibility of the input data. Tips, examples, and pre-defined restricted input options (drop-down menus and radio buttons) minimized the potential for typing errors. Transcribed data could also be exported in Excel format for further scrutiny. O'Shea et al. (2020, 2024) developed a set of classifiers to identify errors in previously transcribed data, analysing records field by field and providing feedback on each entry as well as aggregated accuracy percentages. Despite being a painstaking and time-intensive procedure, this combination yielded high-

quality results, ensuring the accuracy needed to conduct quantitative research on a selection of Dublin and Limerick registration districts. To date, four student transcribathons have been conducted using the HDA for data entry, resulting in 9,022 fully transcribed and cleaned death registration records. These relate to 21 registration districts in Dublin in 1901 and two in 1911, as well as four registration districts in Limerick in 1906. A fifth transcribathon was organised with Clare Roots, a local history society, and 190 records were transcribed relating to the town of Ennis.

4.2 AI TRANSCRIPTION PIPELINE PROJECT

The transcribathons conducted using HDA proved successful in establishing a good quality foundation for a digital, queryable linkage database. However, it soon became evident that manual transcription alone was insufficient to process the extensive volume of data points contained within the dataset. To address this limitation, a parallel approach was developed in the form of the automated, AI-supported Transcription Pipeline Project (O'Shea, 2024). The goal of this project was to enable the automated transcription of the dataset through a collection of specialized Artificial Intelligence (AI) and Machine Learning (ML) models specifically trained for data cleaning, segmentation, and transcription tasks. The resulting pipeline comprises a series of interconnected submodules, each responsible for a distinct task, for example, to make Figure 2 (entry 269, Ellen Walsh) machine readable.

Figure 2 *Register entry 269 Ellen Walsh*

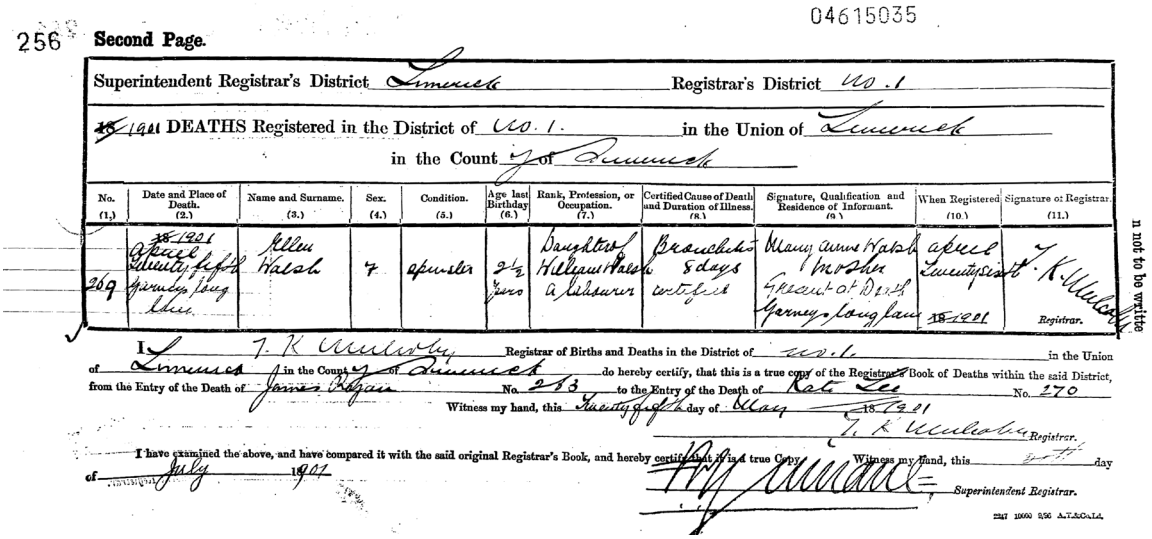


Figure 3 *The high-level workflow of the AI-supported Transcription Pipeline Project*

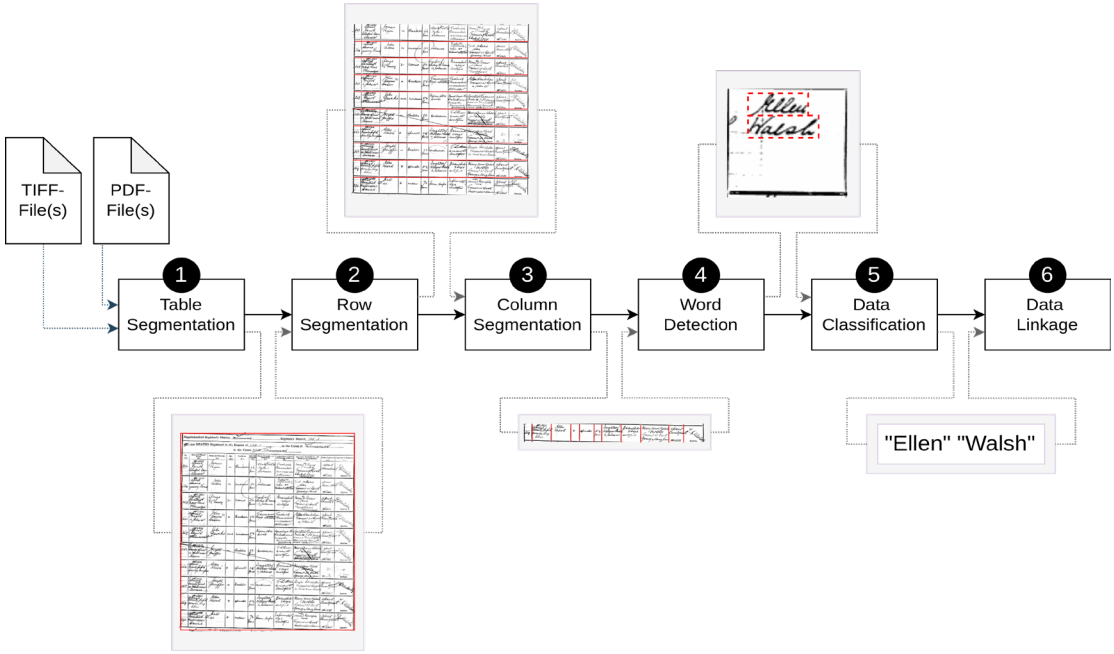


Figure 3 provides a schematic overview of the pipeline. The workflow for the AI-supported transcription pipeline begins with (1) where the source scans are loaded, converted into a uniform format, and cropped to extract the relevant table content. In Row Segmentation (2), the boundaries of individual entry rows are detected and then separated into distinct images. The next stage is Column Segmentation (3). Here the column borders within each row are identified and split into separate cell images. Handwritten text from each cell is extracted during the Word Detection (4) stage. After completion of these first four stages, these elements are classified using a pre-trained AI model in Data Classification (5). Finally, the transcribed data is inserted into a digital database with appropriate linkages established in Data Linkage (6). This workflow was produced and tested on the basis of 2,500 TIFF files, corresponding to ca. 25,000 registered deaths, mostly from the years 1901 and 1911 and from the cities of Dublin and Limerick. Separate work concerns the synthesis data augmentation that took place when increasing the accuracy of the individual phases of the workflow (O'Shea, 2024).

5 WAVE: LEVERAGING LC/NC FOR HISTORICAL CAUSE-OF-DEATH DATA

The Transcription Pipeline described in Section 4 is AI-supported and successfully automates transcription, but primary data users like historians and demographers need programming skills to operate and adapt the workflow to new datasets or different table layouts. This poses a significant skills barrier so since 2024 the computer science goal has shifted to the greater ambition of making the transcription pipeline an autonomous workflow, operating at scale for non-IT competent users.

For historical data transcription to become sustainable and scalable beyond individually funded projects, domain experts must be able to configure and execute workflows independently of IT experts. To bridge this gap, we are developing a comprehensive LC/NC solution using the WAVE platform (https://wave-flow.ie). WAVE's LC/NC paradigm is underpinned by a model centric approach based on the One Thing Approach (OTA) (Margaria & Steffen, 2009). This empowers users to define and evolve workflows either with minimal programming knowledge (Low-Code) or without any coding at all (No-Code). In the context of data transcription, WAVE enables users to model their workflows using predefined building blocks called SIBs (Service-Independent Building Blocks) that encapsulate recurring tasks commonly encountered in the transcription process. The SIB component model was originally standardized by the International Telecommunication Union who sought to establish a vendor and technology independent model which would ensure interoperability for the global telecommunication industry. This model is still used today by the network operators and infrastructure and equipment vendors globally (see https://www.itu.int/rec/T-REC-Q.1200-199709-I/en).

Figure 4 shows how a collection of SIBs may be used to model transcription pipelines for handwritten documents. Several of the depicted SIBs correspond to individual steps from the AI-supported transcription pipeline shown in Figure 3. The blue chevrons at the bottom of each box indicate the SIB's expected data inputs and their types, while the orange chevrons indicate the corresponding data outputs and their types. The SIBs are grouped by functional domain: Normalizers (1) standardize input documents and output data; Segmentation SIBs (2) divide documents into manageable segments; Detection SIBs (3) are used for word and table detection and classification; I/O Operations (4) are for loading documents and creating database entries; and Domain- and Problem-Specific SIBs (5) relate to specialized functions such as trained models or class mappers.

Figure 4 *A SIB palette for modelling transcription pipelines for handwritten documents*

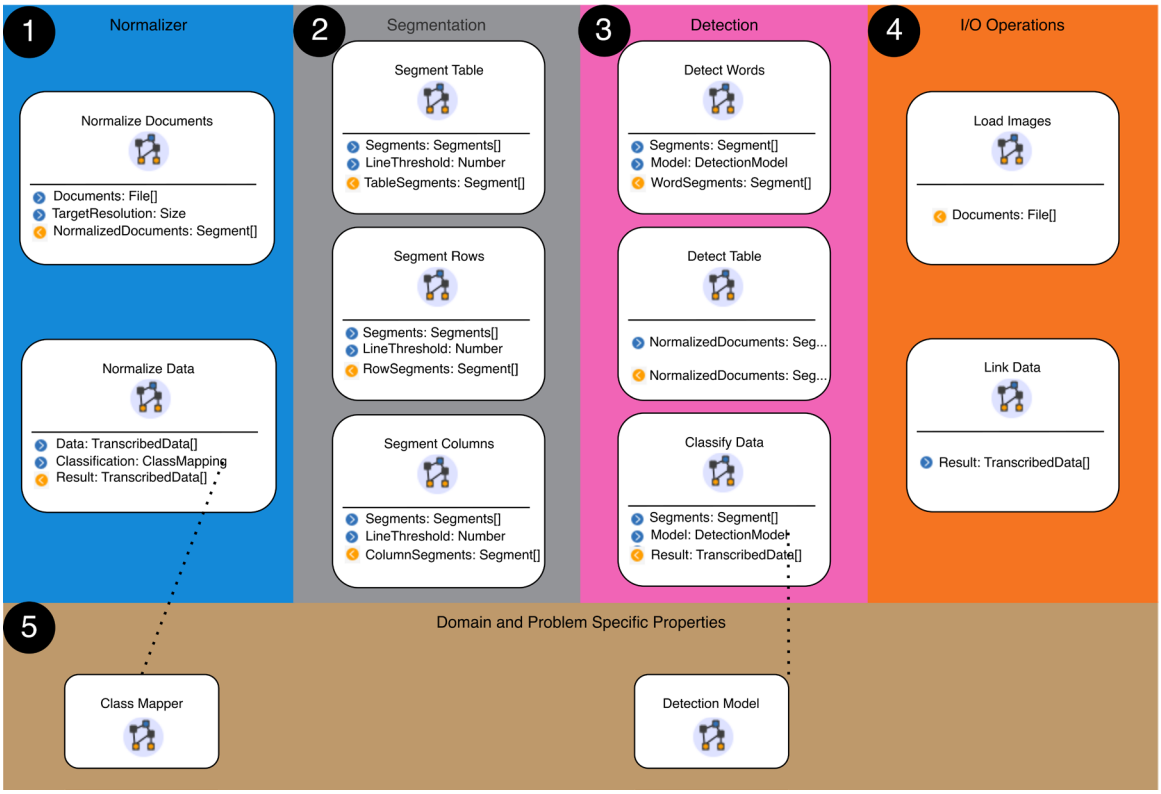
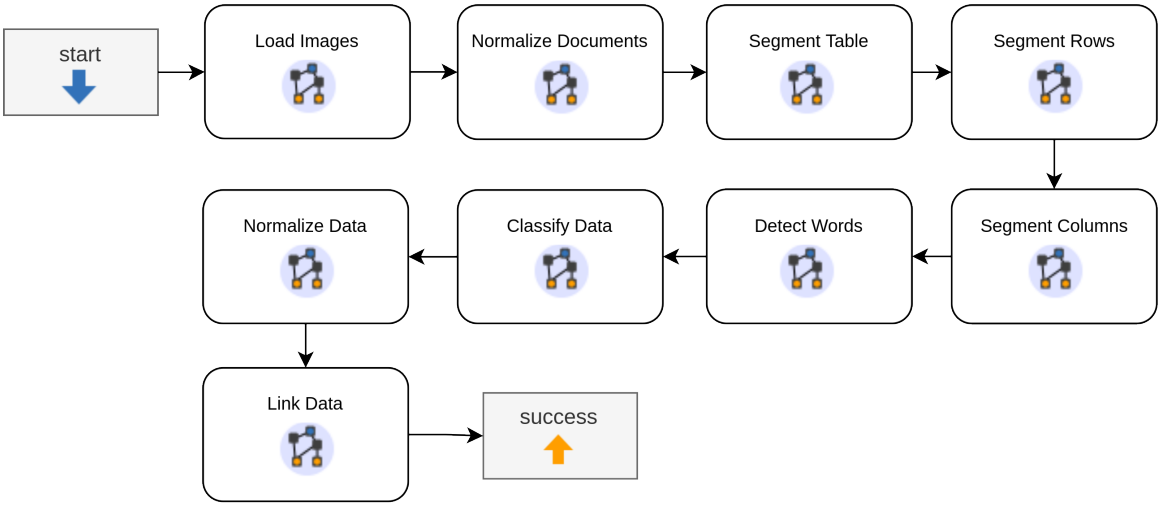


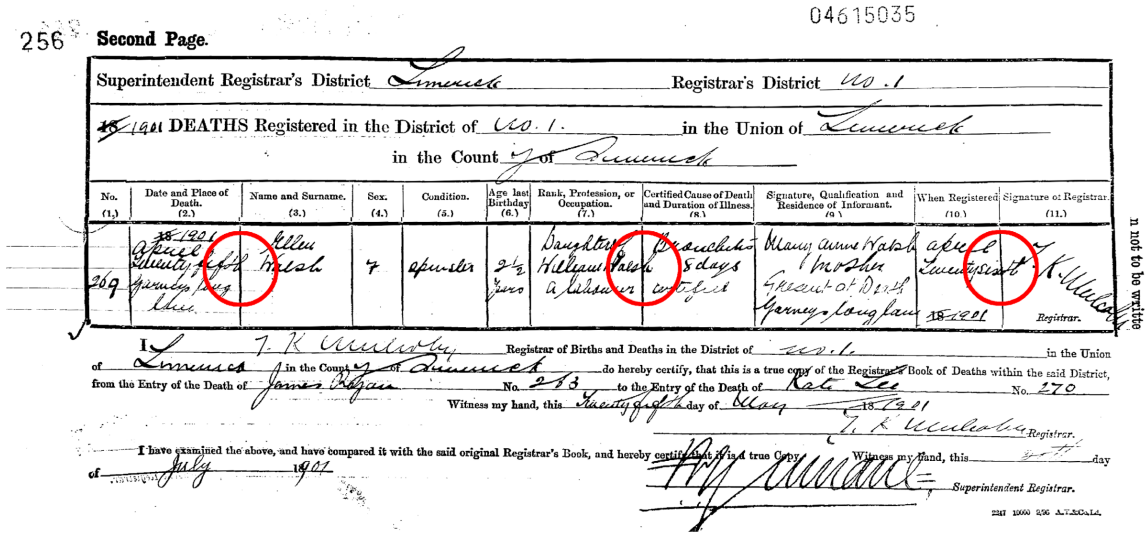
Figure 5 *A simple, linear workflow (control flow only) of the transcription pipeline of Figure 3, as modelled in the WAVE LC/NC platform using the SIB palette of Figure 4*



In this way, the original transcription pipeline is reconstructed using the SIB palette. Figure 5 shows its logical workflow in WAVE's control flow view, with data flow, inputs and outputs intentionally omitted to highlight the similarity between Figures 4 and 6.

The WAVE platform offers several key features that enable easy reconfiguration, correctness, and wide reuse for transcription applications: workflow reconfigurability, conditional branching, parameterization, correctness by construction, and adherence to standards.

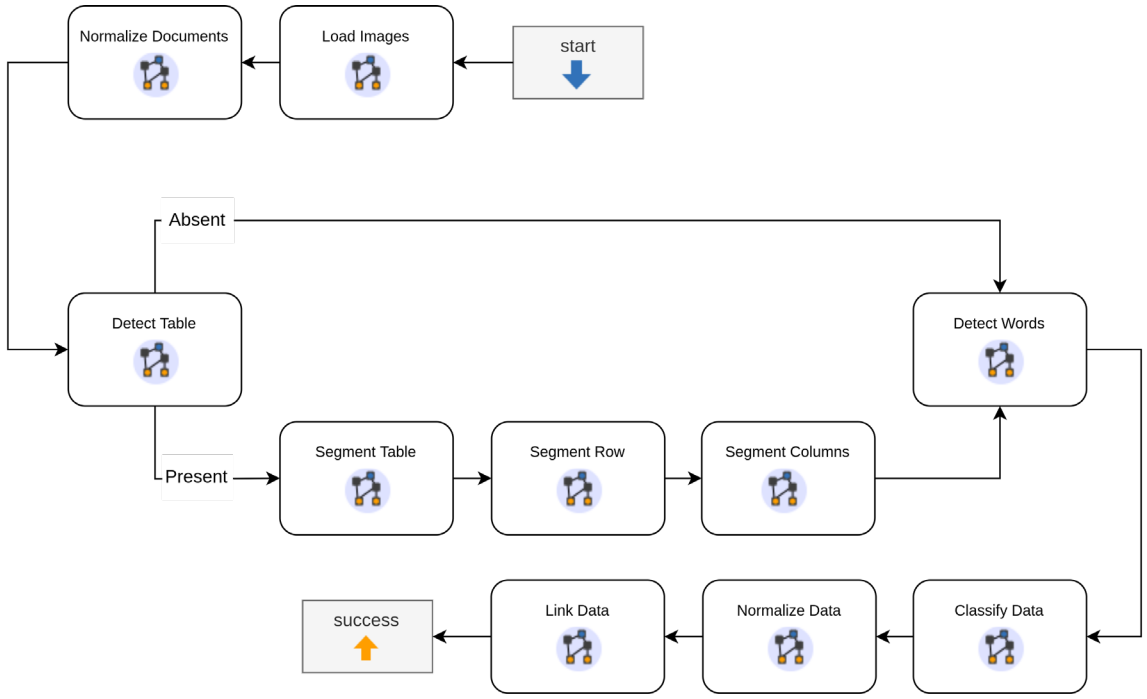
Figure 6 *Overlapping Cell Content being truncated during column segmentation*



Workflow reconfigurability. The platform enables easy reconfiguration of transcription workflows by allowing users to interchange and reorder SIBs in a no-code fashion, tailoring workflows to suit the specific characteristics of each dataset. For instance, Figure 6 shows three points in one row (Figure 2) where handwritten text in one column overlaps into the adjacent column. A straightforward vertical split along the column border would truncate the word, causing misclassification. A domain expert can address this by removing the column segmentation SIB and applying word detection directly to the full row image, a targeted adaptation enabled by the modular SIB palette.

Conditional branching. Datasets often exhibit varying structural characteristics. Figure 7 illustrates a workflow adaptation where the "Detect Table" SIB assesses whether a tabular structure is present. If detected, the workflow proceeds with table segmentation before word detection; if absent, word detection is applied directly. This decision capability enables flexible handling of mixed document types within a single workflow. Note that typical dataflow workflow languages cannot accommodate conditional branching.

Figure 7 *Adaptation of the transcription pipeline showing conditional branching based on the presence of a tabular structure*



Parameterization. Individual SIBs can be fine-tuned through parameters. For instance, the Segment Table SIB requires a "Line Threshold" parameter specifying the expected relative thickness of column border lines, allowing users to adapt the segmentation process to datasets with varying line thicknesses.

Correctness by construction. Although the platform allows free reconfiguration, constraints are enforced to maintain logical consistency. Background sanity checks in the WAVE platform validate workflow structure based on syntax and type consistency, ensuring that workflows created by domain experts execute in a valid and coherent manner.

Adherence to standards. WAVE works with the Common Workflow Language (CWL) standard (https://www.commonwl.org/), enabling interoperability with external workflows such as ICD10h and the Intermediate Data Structure (IDS). Workflows created in WAVE can be shared and reused in other CWL-compatible systems, preventing vendor and technology lock-in and ensuring increased reuse and longevity, which are IT sustainability dimensions.

We envision the WAVE platform as a central, community-driven environment where historians develop, share, and reuse transcription workflows and SIBs. By implementing processing steps once as adaptable SIBs, users can construct task-specific workflow variations tailored to individual datasets, in line with constraint-based modelling (Jörges et al., 2012). SIBs are parameterizable to accommodate diverse document properties such as different scripts, handwriting, languages, or layout formats, for example, a word detection SIB can be configured with an appropriate language model, as demonstrated in a recent case study (El Haji & Margaria, 2025). While the WAVE offers a different prospect for historical demography, much of which is grounded in statistical methods, our aim here is to solve the problem of digitized but non-machine-readable collections. We are committed to the principles of open science, CWL standards and the use of community-driven libraries of curated SIBs. These libraries permit replicability, methodological transparency, and knowledge sharing, which in turn allows researchers to openly share new processing functionality.

Most importantly, the platform aligns with the principles of FAIR (Findable, Accessible, Interoperable, Reusable) research. Each SIB in the curated repository is quality-assessed, versioned, and documented, enabling researchers to trace the provenance of transcriptions, compare methodological variations, and reliably reproduce results.

6 CONCLUSIONS

Scholars of Ireland have made considerable albeit piecemeal headway despite legal restrictions and the general sectoral under-resourcing in cultural heritage and in research on historical demography in higher education. Our work to date has deepened understandings of individual level data, which has established the basis for how the WAVE can ingest and create machine readable data. Further funding is required to create robust research-ready data at a national level: it should be noted that the GRO is not an eligible research performing organisation, but, within a relatively short timeframe we have established a proven track record of results and a strong basis for further collaboration. While our focus in this paper has been on historical cause-of-death data, including birth data is part of our future plans together with the creation of workflows that can enable us to track full life courses at individual levels. For historical demographers and other researchers, ensuring that our civil registration data is available in a machine readable format is essential, along with interoperability between national census and civil registration (GRO) data. Such interoperability will also allow us to understand the way in which communities engaged with their local registrar and census enumerators, and by association how they engaged with the state and various institutions. For instance, it will make it possible to show absence in GRO data in census years. Likewise, fully machine readable birth data is highly desirable as it will allow us to identify mismatches between various datasets, and identify life courses that did not conclude with a death in Ireland. Once we have created these robust datasets, they can be presented in various ways (including for Stata and R), codified using ICD10h, and rendered interoperable with other data types. In contrast with Northern Ireland, England, Wales and Scotland, Irish data are now available free of charge. With linked life course data we aim to examine the Irish records through colonial and post-colonial lens, to ascertain if engagement patterns differed across the four nations, and if there was much divergence in confessional terms from an all-island perspective.

WAVE provides a LC/NC and an open access solution permitting Ireland to play catch-up with other nations mentioned in the introduction, which can conduct more advanced research agendas. In this article we have provided a proof of concept for tracking life courses in countries that, like Ireland, are able to create digitised versions of handwritten originals but have few or no resources to engage in large scale transcription projects. Unlike the widely-acknowledged unreliable nature of AI development and usage patterns, the WAVE platform is founded on the principles of sustainability, which is embedded in the community-driven nature of the approach and the reusability of SIBs (Varoquaux et al., 2025). This creates opportunities for research ecosystems to develop their own workflows while also permitting knowledge sharing and maintenance. For such reasons the resolution of large scale transcription problems for historical demography has utility beyond that domain. The WAVE platform and its development has benefited from the historical cause-of-death use case and it offers a viable option for countries like Ireland, seeking reliable methods to rapidly transcribe handwritten records. It lends itself well to crowd-sourced transcription and the harnessing of voluntary local and family history society efforts to make online resources more accessible.

ACKNOWLEDGMENT

We are grateful for the full cooperation of the Registrar General of Ireland for permission to use these data for research purposes. This research was funded by an Irish Research Council Laureate Award 2017/32 and by Science Foundation Ireland through the grants 13/RC/2094-2 to Lero — the Irish Software Research Centre (www.lero.ie) and 18/CRT/6223 to the Centre for Research Training in Artificial Intelligence.

REFERENCES

Breathnach, C. (2022a). Respiratory disease and death registration. *Annales de Démographie Historique, 143*(1), 39–72. https://doi.org/10.3917/adh.143.0039

Breathnach, C. (2022b). *Ordinary lives, death and social class: Dublin City Coroner's Court, 1876–1902*. Oxford University Press. https://doi.org/10.1093/oso/9780198865780.001.0001

Breathnach, C. (2025). 'The night of the big wind': Reticent subjects and the introduction of old age pensions in Ireland. *History Workshop Journal*, *99*, 29–51. https://doi.org/10.1093/hwj/dbaf002

Breathnach, C., & Gurrin, B. (2017). A tale of two cities — Infant mortality and cause of infant death, Dublin, 1864–1910. *Urban History*, *44*(4), 647–677. https://doi.org/10.1017/S0963926816000717

Breathnach, C., & Gurrin, B. (2018). Maternal mortality, Dublin, 1864–1902. *Social History of Medicine, 31*(1), 79–105. https://doi.org/10.1093/shm/hkx010

Breathnach, C., & Murphy, R. (2023). Fine lines: Locating commercial sex work in official data, Dublin 1901 and 1911. *Urban History, 51*(2), 435–456. https://doi.org/10.1017/s096392682300010x

Breathnach, C., & Murphy, R. (2024). Death and burial data: Ireland 1864–1922 — An interdisciplinary collaboration. In M. Hinchey & B. Steffen (Eds.), *The combined power of research, education, and dissemination. Lecture Notes in Computer Science* (Vol. 15240; pp. 365–376). Springer. https://doi.org/10.1007/978-3-031-73887-6_23

Breathnach, C., Murphy, R., & Margaria, T. (2021). Transcribathons as practice-based learning for historians and computer scientists. *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain* (pp. 1131–1136). IEEE. https://doi.org/10.1109/compsac51774.2021.00155

Breathnach, C., Murphy, R., Schieweck, A., & Margaria, T. (2025). Interoperating civil registration of death and census data: Old age and marriage as categories of analysis. In B. Steffen (Ed.), *Bridging the gap between AI and reality. AISoLA 2023. Lecture Notes in Computer Science* (Vol. 14129; pp. 65–83). Springer. https://doi.org/10.1007/978-3-031-73741-1_5

Breathnach, C., Murphy, R., Schieweck, A., O'Shea, E., Clancy, S., & Margaria, T. (2023). Curating history datasets and training materials as OER: An experience. *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), Torino, Italy* (pp. 1570–1575). IEEE. https://doi.org/10.1109/compsac57700.2023.00242

Breathnach, C., & O'Halpin, E. (2012). Registered "unknown" infant fatalities in Ireland, 1916–32: Gender and power. *Irish Historical Studies, 38*(149), 70–88. https://doi.org/10.1017/s0021121400000638

Breathnach, C., & O'Halpin, E. (2014). Scripting blame: Irish coroners' courts and unnamed infant dead, 1916–32. *Social History, 39*(2), 210–228. https://doi.org/10.1080/03071022.2014.917877

Breathnach, C., & O'Halpin, E. (2021). Sexual assault and fatal violence against women during the Irish War of Independence, 1919–1921: Kate Maher's murder in context. *Medical Humanities, 48*, 94–103. https://doi.org/10.1136/medhum-2021-012178

Brunton, D. (2008). *The politics of vaccination*. University Rochester Press.

Connor, D. S. (2021). In the Name of the Father? Fertility, religion, and child naming in the demographic transition. *Demography, 58*(5), 1793–1815. https://doi.org/10.1215/00703370-9427093

Crawford, E. M. (2003). *Counting the people*. Four Courts Press.

Dahl, C. M., Johansen, D., Sørensen, E. N., Westermann, C. E., & Wittrock, S. (2023). Applications of machine learning in tabular document digitisation. *Historical Methods, 56*(1), 34–48. https://doi.org/10.1080/01615440.2023.2164879

Doherty, A. J., Murphy, R. A., Schieweck, A., Clancy, S., Breathnach, C., & Margaria, T. (2022). CensusIRL: Historical census data preparation with MDD support. *2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan* (pp. 2507–2514). IEEE. https://doi.org/10.1109/bigdata55660.2022.10021106

El Haji, O., & Margaria, T. (2025, November 5). *Democratizing OCR with Low-Code/No-Code image preprocessing* [Talk]. AISoLA 2025, Bridging the gap between AI and reality, Rhodes, Greece. Paperplane.services. https://programme.paperplane.services/event/354

Fernihough, A., Ó Gráda, C., & Walsh, B. M. (2015). Intermarriage in a divided society: Ireland a century ago. *Explorations in Economic History, 56*, 1–14. https://doi.org/10.1016/j.eeh.2014.11.002

Fitzpatrick, D., Gibbon, P., Curtin, C., & Varley, A. (1983). The stem family in Ireland: A debate. *Comparative Studies in Society and History, 25*(2), 339–395. https://doi.org/10.1017/s001041750001046x

Foucault, M. (2003). "*Society must be defended*". *Lectures at the Collège de France, 1975–76* (D. Macey, Trans.). Picador.

Geary, L. M. (2004). *Medicine and charity in Ireland, 1718–1851*. University College Dublin Press.

Gibbon, P., & Curtin, C. (1978). The stem family in Ireland. *Comparative Studies in Society and History, 20*(3), 429–453. https://doi.org/10.1017/s0010417500009075

Guinnane, T. W. (1997). *The vanishing Irish: Households, migration, and the rural economy of Ireland, 1850–1914*. Princeton University Press.

Ireland. Registrar-General, & Farr, W. (1864). *A statistical nosology*. Alexander Thom for H.M.S.O.

Jörges, S., Lamprecht, A.-L., Margaria, T., Schaefer, I., & Steffen, B. (2012). A constraint-based variability modeling framework. *International Journal on Software Tools for Technology Transfer, 14*(5), 511–530. https://doi.org/10.1007/s10009-012-0254-x

Khan, R., Schieweck, A., Breathnach, C., & Margaria, T. (2021). Historical civil registration record transcription using an eXtreme model driven approach. *Proceedings of the Institute for System Programming of RAS, 33*(3), 123–142. https://doi.org/10.15514/ispras-2021-33(3)-10

Kinealy, C. (1997). *A death-dealing famine: The great hunger in Ireland*. Pluto Press.

Margaria, T., & Steffen, B. (2009). Business process modeling in the jABC: The One-Thing Approach. In J. Cardoso & W. van der Aalst (Eds.), *Handbook of research on business process modeling* (pp. 1–26). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-60566-288-6.ch001

Margaria, T., & Steffen, B. (2020). eXtreme Model-Driven Development (XMDD) technologies as a hands-on approach to software development without coding. In A. Tatnall (Ed.), *Encyclopedia of education and information technologies* (pp. 732–750). Springer. https://doi.org/10.1007/978-3-030-10576-1_208

Mourits, R. J., Riswick, T., & Stapel, R. (2024). Common language for accessibility, interoperability, and reusability in historical demography. In B. Steffen (Ed.), *Bridging the gap between AI and reality. AISoLA 2023. Lecture Notes in Computer Science* (Vol. 14129; pp. 10–29). https://doi.org/10.1007/978-3-031-73741-1_2

O'Shea, E. (2024). *Machine transcription pipeline for historical Irish death records* [Doctoral thesis, University of Limerick]. https://doi.org/10.34961/researchrepository-ul.26503576

O'Shea, E., Khan, R., Breathnach, C., & Margaria, T. (2020). Towards automatic data cleansing and classification of valid historical data an incremental approach based on MDD. *2021 IEEE International Conference on Big Data (Big Data)* (pp. 1914–1923). IEEE. https://doi.org/10.1109/bigdata50022.2020.9378148

O'Shea, E., Krumrey, M., Mitwalli, D. S., Teumert, S., & Margaria, T. (2024). From data science to modular workflows changing perspectives from data to platform: DBDIrl 1864–1922 case study. In B. Steffen (Ed.), *Bridging the gap between AI and reality. AISoLA 2023. Lecture Notes in Computer Science* (pp. 84–103). Springer. https://doi.org/10.1007/978-3-031-73741-1_6

Pedersen, B., Islam, M., Kristoffersen, D. T., Bongo, L. A., Garrett, E., Reid, A., & Sommerseth, H. (2024). Coding historical causes of death data with Large Language Models. *ArXiv.* https://doi.org/10.48550/arxiv.2405.07560

Pozzi, L., & Kennedy, L. (2020a). Too long a sacrifice? Maternal mortality in Northern Ireland during the first half of the 20th century. *Annales de Démographie Historique, 139*(1), 141–177. https://doi.org/10.3917/adh.139.0141

Pozzi, L., Kennedy, L., & Manfredini, M. (2020b). Did mothers' lives matter? The protection and promotion of maternal and infant health from the 16th to the 20th century. *Annales de Démographie Historique, 139*(1), 5–26. https://doi.org/10.3917/adh.139.0005

Reid, A., Garrett, E., Dibben, C., & Williamson, L. (2015). 'A confession of ignorance': Deaths from old age and deciphering cause-of-death statistics in Scotland, 1855–1949. *The History of the Family, 20*(3), 320–344. https://doi.org/10.1080/1081602x.2014.1001768

Reid, A., Garrett, E., Hiltunen Maltesdotter, M., & Murkens, M. (2024). *Historic cause of death coding and classification scheme for individual-level causes of death - Manual* [Method]. Apollo - University of Cambridge Repository. https://doi.org/10.17863/CAM.109960

Revuelta-Eugercios, B., Castenbrandt, H., & Løkke, A. (2021). Older rationales and other challenges in handling causes of death in historical individual-level databases: The case of Copenhagen, 1880–1881. *Social History of Medicine, 35*(4), 1116–1139. https://doi.org/10.1093/shm/hkab037

Scalone, F., Pozzi, L., & Kennedy, L. (2023). Religion and child death in Ireland's industrial capital: Belfast 1911. *Social Science History, 47*(3), 425–451. https://doi.org/10.1017/ssh.2023.6

Schieweck, A., Murphy, R., Khan, R., Breathnach, C., & Margaria, T. (2022). Evolution of the historian data entry application: Supporting transcribathons in the digital humanities through MDD. *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), Los Alamitos, CA, USA* (pp. 177–186). IEEE. https://doi.org/10.1109/compsac54236.2022.00033

Varoquaux, G., Luccioni, S., & Whittaker, M. (2025). Hype, sustainability, and the price of the bigger-is-better paradigm in AI. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25). Association for Computing Machinery, New York, NY, USA* (pp. 61–75). https://doi.org/10.1145/3715275.3732006

Acts

Marriages (Ireland) Act, 1844 (7 & 8 Vict.), c.81.
The Poor Relief (Ireland) Act, 1851 (14 & 15 Vict.) c. 68.
An Act for the registration of births and deaths in Ireland, 1863 (26 & 27 Vict. 11), c.11.
Compulsory Vaccination (Ireland) Act (26 & 27 Vict.), c.52.
Births and Deaths Registration Act (Ireland), 1880 (43&44 Vict.), c.18.
Public Health (Ireland) Act, 1878 (41 & 42 Vict.), c.52.

British Parliamentary Papers

First report of Registrar-General of marriages in Ireland. (1850). [1130].
Report of the commissioners appointed to take the census of Ireland, for the year 1841. (1843). [504].
First annual report of the registrar-general of marriages, births, and deaths in Ireland. (1868–69). [4137].
The twenty-eighth report of the deputy keeper of the public records and keeper of the state papers in Ireland. (1896). [C.8163].