Historical Data Preservation and Interpretation Pipeline for Irish Civil Registration Records

Oya Beyan[1], PJ Mealy[1], Dolores Grant[2], Rebecca Grant[2], Natalie Harrower[2], Ciara Breathnach[3], Sandra Collins[2] and Stefan Decker[1]

[1] Insight @ NUIG, National University of Ireland Galway, Galway, Ireland
{ oya.beyan, pj.mealy, stefan.decker }@insight-centre.org
[2] Digital Repository of Ireland, Royal Irish Academy, Dublin, Ireland
{ d.grant, r.grant, n.horrower, s.collins }@ria.ie
[3] Department of History, University of Limerick, Limerick, Ireland
Ciara.Breathnach@ul.ie

Abstract. Semantic Web technologies give us the opportunity to understand today's data-rich society and provide novel means to explore our past. Civil registration records such as birth, death, and marriage registers contain a vast amount of implicit information, which can be revealed by structuring, linking and combining that information with other datasets and bodies of knowledge. In the Irish Record Linkage (IRL) Project 1864-1913, we have developed a data preservation and interpretation pipeline supported by a dedicated semantic architecture. This three-layered pipeline is designed to capture separate concerns from the perspective of multiple disciplines such as archivistics, history and data science. In this study, our aim is to demonstrate best practices in digital archives, while facilitating innovative new methodologies in historical research. The designed pipeline executed with a dataset if 4090 registered Irish death entries from selected areas of south Dublin City.

Keywords: Knowledge Transformation Pipelines . Civil Registration Records . Linked Data. . Digital Archives .

1 Introduction

Semantic Web technologies give us the opportunity to understand today's data-rich society and provide novel means to explore our past. Civil registration records such as birth, death, and marriage registers contain a vast amount of implicit information about society's past, which can be revealed by structuring, linking and combining that information with other datasets and bodies of knowledge. In the Irish Record Linkage 1864-1913 (IRL) project, we adopt Semantic Web and Linked Data technologies to create a platform for storing and linking RDF descriptions of birth, death and marriage (BDM) records for Dublin (1864-1913) to reconstitute families and create longitudinal health histories for the city [1]. The aim of IRL project is to

create a knowledge base, which can serve to answer questions about the accuracy of officially reported maternal mortality and infant mortality rates.

Semantic web and linked data technologies encapsulate the explicit representation of meta-information accompanied by domain theories such as ontologies, which will enable the web to provide a qualitatively new level of services [2]. These technologies have various advantages for capturing and interpreting the civil registration records. RDF metadata enables one to generate different models of data representation for separate concerns or interpretations. Because the linked data is self-describing and explicitly defined in a machine-readable way, it can be linked to external data sets and infer potential relevancies. Also, this feature of linked data supports the capturing of data provenance both as an information source and an interpretation process. By utilizing the rich, expressive features of OWL in a linked data knowledge base, new associations can be discovered. However, knowledge extraction and the presentation process should be sensitive to ethical and privacy concerns such as the preservation of original data, providing a clear description of the interpretation of how the historical data is captured, and not disclosing any personally identifiable information.

In the IRL Project, we have developed a three-layered pipeline to capture, interlink and interpret the civil registration records. In the first layer, we transform the transcribed historical birth, marriage and death records into RDF using the developed Vital Records Ontology (VRO). We enrich those transcriptions by creating links between partial graphs with the Historical Events Ontology (HEO) and annotate relevant data with domain coding standards such as cause of deaths in the second layer. The third layer captures the defined sets of interpretations for a given competency question and returns aggregated de-identified information to the domain experts. The developed transparent data preservation and interpretation pipeline is supported by a dedicated semantic architecture project and adheres to Linked Data principles. In this study, our aim is to demonstrate best practices in digital archives while facilitating innovative new methodologies in historical research.

1.1 Motivation

Our motivation is to develop novel methods to explore and interpret historical data sets with semantic web technologies and Linked Data. Digital repositories provide a central access point and interactive tools for historical and contemporary data [3]. These repositories may serve diverse interest groups such as archivists, historians, journalists, public researchers and scholars. The developed knowledge infrastructure should satisfy different, sometimes conflicting perspectives and concerns, as well as support privacy of the data subjects.

In this study, we have developed three layers for storing, exploring and interpreting these Irish civil registration records. We demonstrate our concept with

the infant deaths as the use case. Infant death is an important social and political indicator of human welfare and national wealth and social conditions such as poverty and single motherhood [4,5]. The pipeline will initially include 444 death register pages, which equated to 4090 death entries recorded in two Registrar Districts to the South of Dublin City from the years 1870 and 1890.

2 Methods – writing in progress

During the period 1864-1913, birth, death and marriage records were manually entered on a register by a Registrar and the true copy was later verified by a Superintendent Registrar. The project data consists of digitised birth, death and marriage register pages as well as a corresponding database, shared for the duration of the project by the General Register Office of Ireland. As it is presently a closed dataset, access to this data is restricted to IRL team members and no persons can be identified from the project outputs. Each register page may include up to 10 records, each one registering the birth, death or marriage of an individual. The digitised records were analysed by the digital archivists and broken down to identify all information captured in a given record and register page. It was necessary to curate a new database for the purposes of the project, in which the digital archivists transcribe a sample of the birth, death and marriage records from the Registrar's Districts of Dublin South City 1 and Dublin South City 3. Death records have been focused on initially, as the historical research questions focus on infant and maternal mortality. Using the original database for reference, as well as manual curation, relevant records were selected and transcribed in the new database to capture all original information. The register page and the records thereon were linked to preserve the original context of record creation.

3 Data Preservation and Interpretation Pipeline

In the IRL project, we have developed a three-layered pipeline to capture, enrich and allow for new interpretations of the historical data.

The aim of the first layer is to preserve the civil registers in their original form and capture the provenance of the archival record. From the digital archivist's point of view, the register pages are the main units to be preserved. The VRO is developed to annotate each register page and preserve the authenticity. In this layer, we converted the historical data into Linked Data and preserved them in the original order and without any interpretation.

The second layer is dedicated to creating links between the captured records and identifying the associations between them, for instance, using nominal and geographic data, individuals and familial bonds can be identified and subsequently verified by address. It also includes annotations to other standards or ontologies

such as the cause of deaths. The HEO was developed to enrich the registers and interlink each archival entry to constitute families.

The final layer is designed for exploring the linked records stored in the second layer from various points of interest. In this layer the data is queried which permits historians to examine the de-identified results from several perspectives. For example, the definition of maternal mortality is historically poorly defined but the pipeline permits historians to reinterpret the data in order to potentially identify additional deaths [6]. This layer permits researchers to apply different definitions, for instance the WHO's current definition of a direct maternal death is one occurring within 42 days of the delivery or termination of pregnancy [7]. Use case specific ontologies can enable the historical data to withstand multi-factorial queries for example, timeframes for deaths from puerperal sepsis (a common cause of maternal death) can be cross-referenced with the ages of the women involved to reveal patterns in maternal mortality.

In the following sections, we will describe the role of each layer in detail and present concrete examples.

3.1 Preservation Layer

The first layer serves as a long-term digital preservation platform for digitised objects, namely Register Pages for this specific study. . Register pages are transcribed verbatim in the original form and represented in Linked Data format. The aim of this layer is to provide a trustworthy platform for preserving the historical data by applying digital archival principles. Linked data structures are designed based on the provenance and archival authenticity principles.

Archival theory is based on two key principles, respect de fonds (original order) and archival provenance. Respect de fonds is the principle which guides archivists when exerting intellectual control over a collection, and ensures that the archival record is always described in relation to the context in which it is created as far as possible. We follow this principle by transcribing not a line of data about an individual, which is meaningless in an archival context, but the entire register page that constitutes an archival record or object. Provenance refers to how the archival record relates to its creator, and can only be maintained through the appropriate description of an archival record. VRO is designed to fulfil these two basic principles.

VRO has two basic classes for representing a digital object and its data, namely RegisterPage and Record. Births, deaths, and marriages were captured per district (within a union, within a county) as single records on register pages. Each page can contain up to 10 records, after which it is signed off by the registrar and sent to the superintendent registrar for inspection and validation. The RegisterPage object encapsulates the metadata of the physical Register Page including dates, place, volume, stamp number as a unique identifier, District Registrar, and Superintendent

Registrar. The Record object captures data in the Register Page with exactly same attributes such as name, forename, and date of birth. Because one of the projects aims is to maintain the original record by minimizing interpretation, we choose to develop a "flat" ontology, which means that most information that can be found on such a register page was captured as literals. A RegisterPage and a Record are linked; each record must belong to a register page and each register page can have zero or more records.

3.2 Interlinking and Enrichment Layer

The aim of the interlinking and enrichment layer is to facilitate the exploitation of historical data for various purposes by enabling efficient queries. The data schema of the VRO were designed for preserving digitised objects as they are. Therefore, the triple storage created with the VRO ontology is not particularly effective for exploring the data through generations and gaining insights into longitudinal health histories.

The interlinking creates associations between different pieces of information captured from register pages and allows for the reconstitutes of families from the data captured in historical events. This requires the interpretation of the historical data at varying levels. The Historical Event Ontology (HEO) is developed to represent the structure of the historical event in terms of actors, events and their relations with each other. Metadata for each level of interpretation is held with the HEO and linked to the original record to follow the provenance.

In the first level of interpretation, we identified the actors of birth, death and marriage events as they are represented in the civil registration records. Depending on the historical event there are a different number of actors participating in each record. For example in a death event, four different people are identified, i.e. i.e. the person who has died, the informant, the District Registrar and the Superintendent Registrar. In the next step, the actors are linked with each other according to the role they played in the historical event. Figure 1 represents the relations between the actors and events.

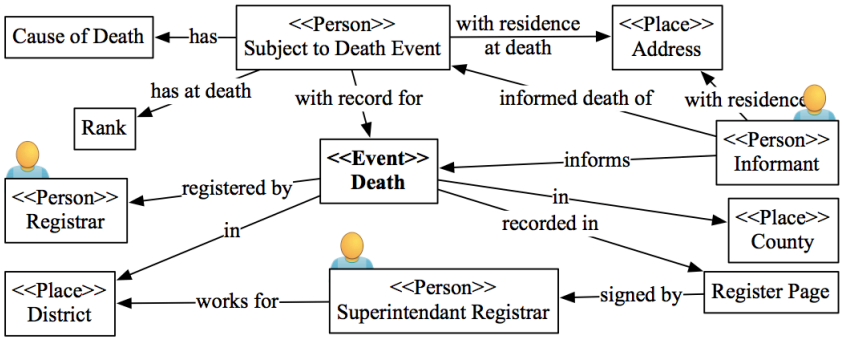

Fig. 1. Interpretation of the actors and their relations for a death record with the HEO.

The first level of interpretation is based on the structural attributes of register page and is more straightforward. In the second level of interpretation, we have investigated data fields to identify relations between actors. For example the informant of a death event might be the father or wife of the deceased person. The relationship between the informant and the subject of a death certificate is captured as a data attribute of the RegisterPage object. With keyword search, relevant entries are linked with the object properties of the HEO. To look at another example, the same person might participate in more than one event and therefore, is created more than once. As figure 2 illustrates, Mary Jane might have a marriage record and the same Mary Jane might appear in a birth record as the mother of a child. Then these two linked data URIs are interlinked with the owl: SameAs annotation.

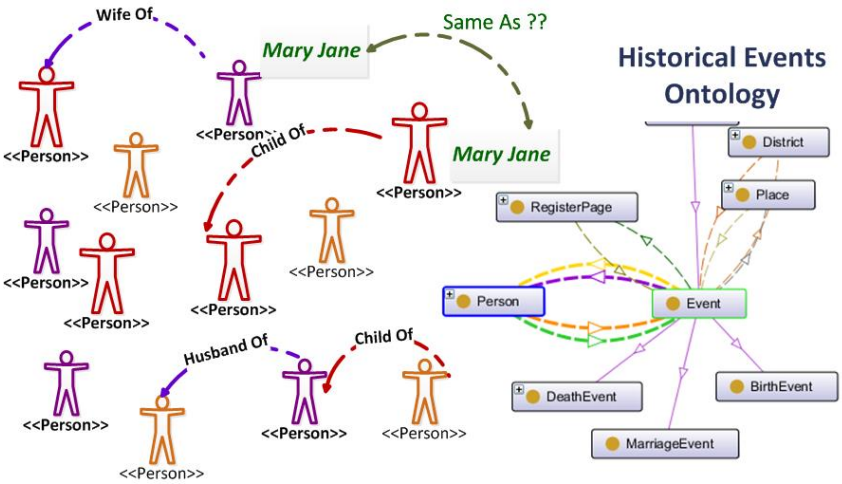

Fig. 2. Interlinking URIs with the HEO object properties.

The third level of interpretation is to enrich the existing data set with standard terminologies and ontologies. Attributes such as place name and cause of deaths can be annotated with related nomenclatures and coding systems. In this study, we examined the cause of death and mapped them to different coding systems. Medical coding systems evolve over time. In 1864 all Irish Registrars were furnished with copies of a standard nosology, which identified 145 causes of death [8]. Reflecting significant advances in medical science medical coding systems underwent a similar evolution in the period under review 1864-1913. Using the causes of death in the 1890 sample as a guide we explored the coding systems used in that time frame. To supplement the 1864 nosology we selected three available coding systems namely, the International List of Causes of Death, Revision 1 (1900) (ILCD1), the International List of Causes of Death, Revision 2 (1909) (ILCD2), and the International Classification of Causes of Sickness and Death (ICSD) [10,11,12]. The distinct cause of deaths is selected from the triple store, manually reviewed by the domain experts, and mapped to the available codes in ILCD1, ILCD2, and ICSD. In HEO, we created CauseOfDeath and identified subcategories for each of them. Each subcategory is annotated with the relevant ILCD1, ILCD2 and ICSD codes. As shown in figure 3, in the linked data repository a person object is linked with a blank node, which contains the original cause of death and duration of illness. Then individual causes of death are classified with the defined CauseOfDeath subcategories.

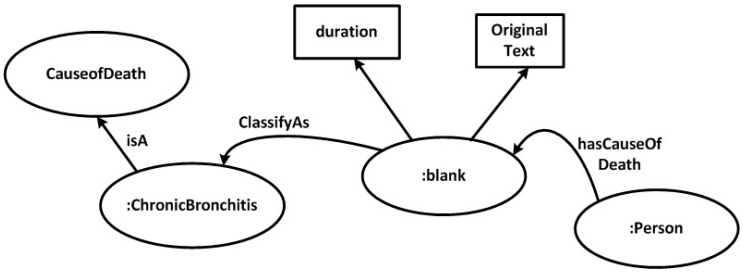

Fig. 3. Enriching the death records with ILCD1, ILCD2, and standards.

In this layer of the semantic pipeline, we interlink data to reconstitute families and enrich data with coding systems. The HEO enabled us to transform linked data records created with the flat VRO ontology to graphs of people, events, and relations. This new structure enables researchers to effectively query the linked data repository and access even more information about the past life event. However, it makes the linked data repository more fragile to privacy violations. Therefore, we implemented a third layer to explore the data with defined use cases which will return aggregated statistical data rather than the individual level of details.

3.3 Use Case Query Layer

The ultimate aim of the semantic pipeline is to provide historians with tools to analyse historical events and to answer their specific research questions such as "How accurate are historic maternal mortality rates and infant mortality rates for Dublin?" Historic definitions vary for maternal and infant mortality. Maternal death is defined as the death of a woman while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management but not from accidental or incidental causes [9]. Infant mortality is currently defined as a death of a child born in a specific year or period dying before reaching the age of one, if subject to age-specific mortality rates of that period. Deaths in the first 24 hours or in the first 27 days also have specific significance from the historians' perspective.

The use case query layer enables researchers to set their questions and define varying versions of concepts they are interested in. In the infant mortality use case, infant mortality is examined from multiple perspectives including the time frame of death, seasonality, location and the cause of death. Death time frame is defined with four classes; *deathIn24hours*, *deathIn27days*, *infantDeath*, and *neoNatalDeath*. Results of queries are returned in aggregated form without disclosing any identifiable personal data. The death timeframes correspond with specific disease and whether or not the infant was weaned, which is indicative of lower socio-economic circumstances.

4 Implementation

We have implemented the proposed pipeline with linked data standards and serve over the JENA Fuseki SPARQL endpoint. Figure 4 shows the implantation steps of the pipeline. During the data acquisition phase, the digital archivists transcribed the register page information and the individual birth, death and marriage records into a MySQL database.

In the preservation layer, D2RQ Mapping is applied to extract the data from the MySQL database into RDF using the VRO. D2RQ is a system used to relational databases as virtual, read-only RDF graphs. It allows for the creation of custom dumps of the database in RDF formats for loading into an RDF store [13]. In the RDF representation, we have utilized the VRO. In the mappings, special care was taken to preserve the ability to trace information back to the source- the original records. The transcriptions included the original page numbers and unique register stamp numbers, as well as the name of the Registrar and Superindenent Registrar.

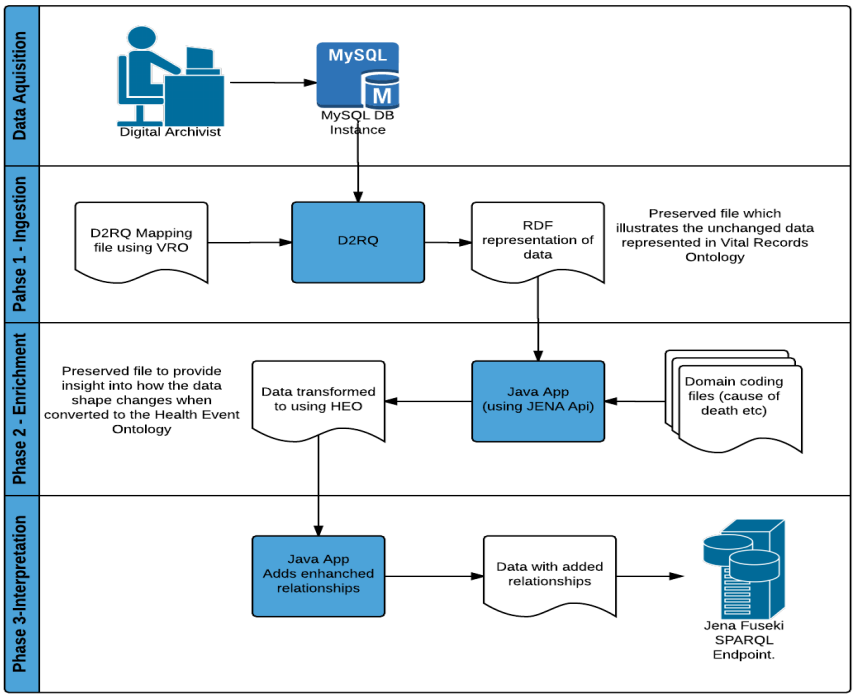

Fig. 4. Implementation of semantic pipeline .

The interlinking and enrichment layer was implemented with a Java application using JENA. VRO based linked records were processed and converted to the HEO based linked data. The Java app makes use of the Apache Jena API to load the data in RDF form from the output of the D2RQ mapping. It works through each record, (death, birth, marriage) and develops a new data model based on the HEO. The resulting linked data contains the object classes detailed in the HEO ontology such as different types of Person (Registrar, Informant, Superintendent) and the various Event types (Death, Birth, Marriage)

During this process, the HEO records are progressively enhanced to add linkages to allow for identification of individuals and to carry out normalizations such as aligning causes of death with ILCD standards. Historical causes of death, which were recorded in the death records, often include medical terminology which by today's standards may appear cryptic or inconsistent. The Java app loads a custom file which contains mappings for the domain of causes of death (as found in the data) to a standardized set of international causes of death. The correct mapping for each record is identified and added to the data set.

Another example of these kinds of enhancements is the addition of an ageAtDeathInMinutes field. The original death records contain may text in the 'age at death' field such as "about 60 years" or "2 years and 3 months". In this form, the data would not lend itself to querying very well, for example, or would not

effectively identify children who died under the age of 2. The application code, therefore, takes the textual representation of age at death and converts it to a numerical minutes field.

In the final phase, JENA Fuseki SPARQL endpoint serves to answer the use cases and return the queries.

5 Discussion and Future Work

Semantic technologies and Linked Data promises many advantage for capturing, exploring and interpreting the historical data sets. Ontologies provided means for separating varying concern, preserving authenticity and following the provenance of the records.

Acknowledgements We thank the Registrar General of Ireland for permitting us to use this rich digital content contained in the vital records for the purposes of this research project. This publication has emanated from research conducted within the Irish Record Linkage, 1864-1913 project supported by the RPG2013-3; Irish Research Council Interdisciplinary Research Project Grant, and within the Science Foundation Ireland Funded Insight Research Centre (SFI/12/RC/2289). The Digital Repository of Ireland (formerly NAVR) gratefully acknowledges funding from the Irish HEA PRTLI programme.

References

[1] Beyan, O., Breathnach, C., Collins, S., Debruyne, C., Decker, S., Grant, D., ... & Gurrin, B. (2014, July). Towards Linked Vital Registration Data for Reconstituting Families and Creating Longitudinal Health Histories. InKnowledge Representation for Health Care KR4HC 2014. Organized under the" Vienna Summer of Logic 2014" multi-conference.

[2] Davies, J., Fensel, D., & Van Harmelen, F. (Eds.). (2003). Towards the semantic web: ontology-driven knowledge management. John Wiley & Sons.

[3] N. Harrower, S. Webb, J. Tang, D. Gallagher, E. Kilfeather, S. O'Tuairisg, S. Collins. (2013) Developing the Irish National Trusted Digital Repository for the Humanities and Social Sciences: an interdisciplinary approach. OR2013.

[4]Breathnach, C., & O'Halpin, E. (2012). Registered 'unknown'infant fatalities in Ireland, 1916–32: gender and power. Irish Historical Studies, 38(149), 70-88.

[5] Breathnach, C., & O'Halpin, E. (2014). 'Scripting blame: Irish coroners' courts and unnamed infant dead, 1916–32'. Social History, 39:2, 210-228, DOI:10.1080/03071022.2014.917877.

[6] Rebecca Kippen, 'Counting nineteenth-century maternal deaths: the case of Tasmania', Historical Methods: A Journal of Quantitative and Interdisciplinary History, 38:1 (2005), 14-25.

[7] WHO. International Classification of Diseases and Related Health Problems (Geneva: World Health Organization, 1992).

[8] Registration of deaths in Ireland: a statistical nosology, comprising the causes of death, classified and alphabetically arranged with notes and observations (Dublin, 1864).

[9] WHO Health Statistics and Information Systems,
http://www.who.int/healthinfo/statistics/indmaternalmortality/en/ [July 15, 2015]
[10] International List of Causes of Death, Revision 1 (1900)).
http://www.wolfbane.com/icd/icd1h.htm [July 15, 2015]
[11]International List of Causes of Death, Revision 2 (1909)
http://www.wolfbane.com/icd/icd2h.htm [July 15, 2015]
[12] Department of Commerce and Labor, Bureau of Census (1910). International Classification of Causes of Sickness and Death. Washington Government of Printing Office.
[13] Bizer, Christian, and Richard Cyganiak. "D2r server-publishing relational databases on the semantic web." Poster at the 5th International Semantic Web Conference. 2006.