Towards Linked Vital Registration Data for Reconstituting Families and Creating Longitudinal Health Histories

Oya Beyan[1], Ciara Breathnach[2], Sandra Collins[3], Christophe Debruyne[1,3], Stefan Decker[1], Dolores Grant[3], Rebecca Grant[3], and Brian Gurrin[2]

[1] Insight @ NUIG, National University of Ireland Galway, Galway, Ireland
{firstname.lastname}@insight-centre.org
[2] Department of History, University of Limerick, Limerick, Ireland
{firstname.lastname}@ul.ie
[3] Digital Repository of Ireland, Royal Irish Academy, Dublin, Ireland
{s.collins,c.debruyne,d.grant,r.grant}@ria.ie

Abstract. The Irish Record Linkage 1864-1913 project aims to create a knowledge base containing historical birth-, marriage- and death records encoded into RDF to reconstitute families and create longitudinal health histories. The goal is to interlink the different persons across these records as well as with supplementary datasets that provide additional context. With the help of knowledge engineers who will create the ontologies and set up the platform and the digital archivist who will curate, ingest and maintain the RDF, the historians will be able to analyse reconstructed "virtual" families of Dublin in the 19th and early 20th centuries, allowing them to address questions about the accuracy of officially reported maternal mortality and infant mortality rates. In the longer term, this platform will allow researchers to investigate how official historical datasets can contribute to modern-day epidemiological planning.

Keywords: birth, death and marriage records; health histories; ontologies and linked data; automatic annotation.

1 The Introduction and Problem Statement

To date, the history of medicine and epidemiology in Ireland has focused on medical education and disease administration; the body of the patient remains under-explored [5], [7]. In historiographical terms Irish infant and maternal mortality rates have not received due scholarly consideration. This is due to the limited accessibility of vital registration data. Underreporting of births and deaths, for cultural reasons, is a problem in Ireland in the era we are dealing with and currently in India and Sub-Saharan Africa – not to mention the closed societies of the Far east. The ramifications for local and central governance are huge from a public health planning perspective but on a global scale, the WHO has identified Vital Registration under-reporting as a major obstacle to improvements in MMR and IMR and of course disease control. Through our micro study we

hope to find adjustments to official figures, which we hope can have modern day applications. Building on the seminal work of Geary [5] and Jones [7], we aim to determine how a granular vital records dataset can be useful to modern-day epidemiology and public health strategies.

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. This will allow us to address research questions such as: "How accurate are historic maternal mortality rates (MMR) and infant mortality rates (IMR) for Dublin?" Historic definitions vary for MMR and IMR. Maternal death is defined as *deaths of women while pregnant or within 42 days of the end of the pregnancy from any cause related to or aggravated by the pregnancy or its management, but not from accidental or incidental causes* [1]. This definition is based on the International Statistical Classification of Diseases and Related Health Problems (ICD).[1] Infant mortality is currently defined as a death under 1 year.[2]

2 Approach

The creation and use of the IRL platform involves three types of "active participants": the historian, aiming to analyse the information stored in vital records for his own research agenda, the digital archivist, responsible for archiving and digitally curating the vital records, and the knowledge engineer, who is responsible for creating an appropriate ontology and setting up the Linked Data infrastructure. Fig. 1 provides a graphical representation of the IRL platform. The digital archivist fulfils the role of both a traditional archivist and a digital data curator. The digital archivist's tasks include preparing the digitised data from the General Register Office (GRO) records for transformation into RDF, quality control of the linked datasets, and contributing to the selection of additional external datasets to add contextual information. The digital archivist will also ensure that best practice in data publication, data protection and digital preservation are adhered to during the project.

The RDF created by the digital archivist's input as well as the selection of additional datasets from the Linked Open Data (LOD) cloud – in this figure depicted by a rescaled LOD cloud diagram[3] – will be given as input to a "linker" that will assert entity equivalences based on a predefined set of rules using the Silk framework.[4] Using these rules, the Silk framework compares entities from the different datasets and assigns values for pairs of entities indicating their similarity. The higher that value, the more likely they are to be equivalent according to the Silk framework. The discovered links as well as updates from the digital archivist are used to update the triplestore. The triplestore is accessible via a

[1] `http://apps.who.int/classifications/icd10/browse/2010/en`
[2] `http://www.who.int/whosis/whostat2006DefinitionsAndMetadata.pdf?ua=1`
[3] LOD cloud diagram by R. Cyganiak & A. Jentzsch. `http://lod-cloud.net/`
[4] `http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/`

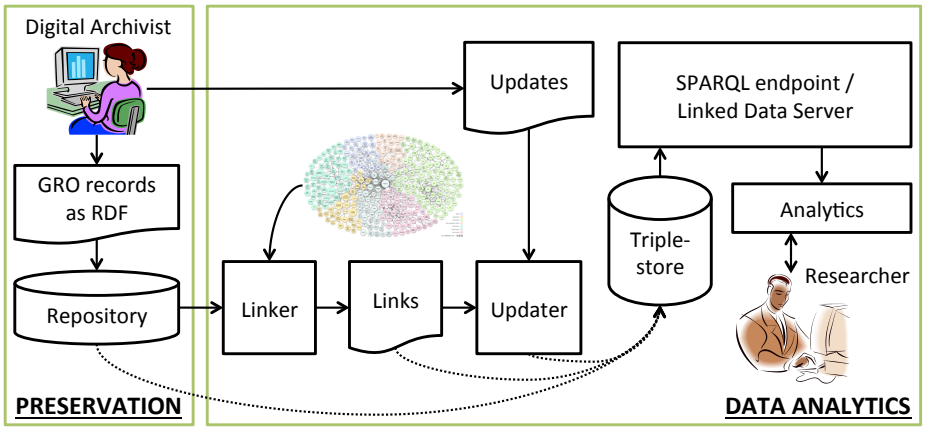

Fig. 1. Conceptual architecture of the IRL platform.

SPARQL endpoint and a Linked Data server, on top of which various tools can be built to analyse the digitised GRO records.

It is important to note that the terms and conditions of our data sharing agreement do not permit us to make public any data that would identify any individual person. Under Irish law great care is taken to ensure that access to public records protects individual rights to privacy.[5] It is permissible to access the historic records of the GRO at its dedicated research room in Dublin, but it is restricted per diem and there is an associated charge. This data is also available on commercial sites such as `ancestry.com` and `familysearch.org` as individual name searches. In the UK, for instance, census records enter the public domain after 100 years.[6] This convention has also influenced the parameters of this study. At this moment, we choose to restrict access to the RDF, SPARQL endpoint and Linked Data Server. Notwithstanding the fact one can also argue whether public domain data needs to be anonymised, we anonymise all persons in the RDF representations of the GRO records.

IRL will apply linked data technologies to birth, death and marriage (BDM) records (1864-1913) to reconstitute families and create longitudinal health histories. The initial study will concentrate on Dublin city, for which the records are reasonably robust. The records contain information posing quite a few challenges. Personal names, place names and ages were inconsistently recorded. Age presents a particular challenge, as, in many cases, people were not aware of their true age; often births of children were registered only after parents knew they were going to survive. Appropriate methods are needed to interlink the different datasets and will help us to reconcile matters.

[5] `http://www.nationalarchives.ie/PDF/NAA1986.pdf`
[6] Lord Chancellor's Instrument no.12, issued in 1966 under S.5 (1) of the Public Records Act 1958.

3 Ontology Construction

Two types of ontologies are created for the IRL platform: an ontology for the creation of RDF descriptions of the GRO records' contents, and ontologies for supporting the researcher's data analysis tasks. The first ontology is created by looking at the data contained in GRO records. The digital archivist will ensure an exact encoding of the content in these records into RDF, stored in an appropriate repository. (see Section 4). The creation of the second type of ontology is more challenging. Given that the stakeholders wishing to analyse the content – in this particular project historians – are not necessarily familiar with ontology engineering and the knowledge base needs to support their activities, we adopted the approach proposed by Grüninger and Fox of having the stakeholders formulating *competency questions* [6]. The ontology must contain a necessary and sufficient set of axioms to represent and solve these questions [6]. These competency questions are not used to generate an ontology, but rather to evaluate it [4]. Using the types of queries the stakeholders wish to see answered, the knowledge engineers built an ontology, which was specifically tailored for the project, yet aimed to reuse existing, established vocabularies where possible.

Competency questions formulated by historians include: "How many women died within n days after childbirth due to complications related to labour and how does that figure correspond with the official reports?", "What women died of causes that can be attributed to maternal death, but for which no corresponding birth certificate exists?", "What is the average sibship interval?", "How did various socio-economic conditions (e.g., before and after marriage, the profession or trade of the father, etc.) affect maternal and infant mortality rates?", "What is the average sibship interval where the first child did not survive under various socio-economic conditions?" and "What are the MMR and IMR across counties and parishes?"

The questions were analysed to identify the concepts and relations for the ontology, which were validated by the stakeholders. Ontologies were discussed by means of graphical representations, e.g., as shown in Fig. 2. This figure shows some of the concepts and relations one can find in birth records. Note that some relations were captured as concepts – such as residence, marriage, and baptism – to support the competency questions. To implement the ontology, we looked at existing ontologies for reuse and integration as well as the creation of missing concepts and relations to meet the requirements. FOAF (Friend-of-a-Friend) RDF schema provides concepts related to the representation of personal information and social relationships. The main aim of the FOAF project is to find information about people, and it focuses on constructing links between people [3]. The Persona Data Model[7] aims to control how personal data is shared within social networks and uses the Persona Vocabulary[8] providing definition for personal attributes. Both of these efforts aim to collect and share personal information and relations that exist in today's Semantic Web environment. However,

[7] `http://wiki.eclipse.org/Persona_Data_Model_2.0`
[8] `http://wiki.eclipse.org/Persona_vocabulary`

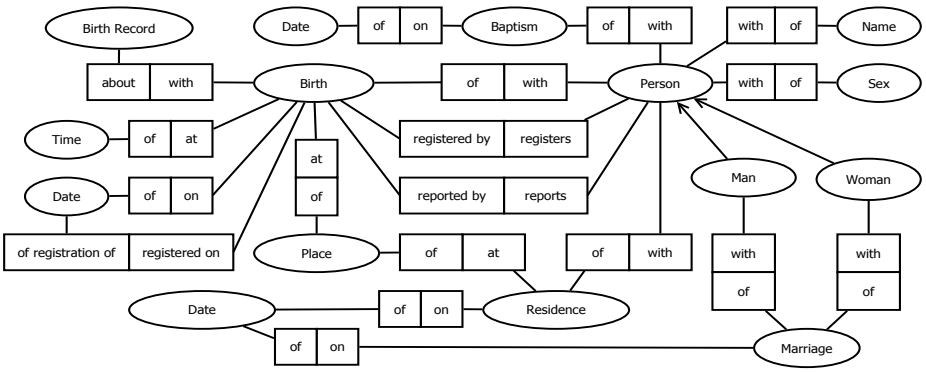

Fig. 2. Graphical representation of concepts and relations identified in birth records.

our research targets historical reconstitution of relationships between persons, their roles played in families and their health histories. Therefore, we have to deal with concepts related to time dimensions and roles. For the second type of ontologies, we also reused available domain disease ontologies [2], personal information vocabularies, and we developed our novel ontology for dealing with defined domain problems, e.g. by formalising information found in classification systems such as ICD.

Instances of concepts and relations of the second ontology will partly stem from the "flat" RDF representations of GRO records by applying mappings, transformation and heuristics. The additional triples will be stored separately, as to ensure a clear separation of concerns (preservation and reusability versus data analysis).

4 Data Annotation and Ingestion

The Digital Repository of Ireland (DRI) is an interactive, trusted digital repository for social and cultural content held by Irish institutions. As a national digital infrastructure, the DRI works with a wide range of institutional stakeholders to link together and preserve Ireland's rich and varied humanities and social science data. Based at the DRI, the IRL-DRI digital archivist applies best practice standards in data curation and digital preservation to ensure that the authenticity of the archival record is maintained in order to be reliably reused. The original GRO dataset has been exported from a relational database and prepared for transformation into RDF to create the new Linked Data platform. Data preparation and curation is undertaken following best practice in digital archiving to facilitate content searchability, discoverability and overall data management. An essential requirement for the construction of the Linked Data platform is that each entity is given a unique identity, in the form of a Unique Resource Identifier (URI). As well as serving the purpose of identifying and expressing the objects, the assignment of a URI removes any ambiguity between

people of the same name; a key concern in relation to vital registration data from 19th and early 20th century Ireland. The creation of a specialist metadata vocabulary for the description of the archaic terms found in historical records aids digital retrieval and improves consistency in the dataset.The digital archivist's role in the project enables data use without distortion through quality control of the linked dataset and the ingestion of the (into) RDF encoded GRO records into the DRI repository that supports RDF by building on top of the Fedora Repository Project.[9]

5 Links with External Datasets and Data Analytics

The goal of the IRL platform is to use and discover contextual information in and from other datasets. The contextual information is either discovered and entered by the digital archivist, or discovered with the Linker or data analytics component of the IRL platform. A number of potential datasets have been identified which could be used to add contextual information, depending on their terms of access and reuse: (i) the Irish Genealogy platform, which has parish records for certain counties[10]; (ii) the Online Historical Population Reports, which contains legacy datasets from other funded research projects[11]; (iii) census data for 1901 and 1911 from the National Archives of Ireland[12]; and (iv) Logainm[13], which contains information on Irish place names, and even some street level information for Dublin. The latter even stored this information as RDF, linked with DBPedia[14] and accessible via SPARQL.

We now give one example on how we aim to use contextual information to classify certain deaths as maternal mortality. Deaths were registered by an informant with the local Registrar, who entered them in registration volumes. These volumes were later forwarded to the Superintendent Registrar. A death record includes individual identification information; register number and page numbers. Certified cause of death and duration of illness, place of death, date of death and personal information about the deceased such as name, sex, marital status, age and profession are also recorded. The informant is also identified in the record as are the registrars and superintendant registrars. While the entry fields remain the same over the period 1864-1913, the level of detail inputted by the Registrar in each field can vary significantly. An example of a cause of death is "hemorrhage", which is not immediately related to maternal death. But, if a woman died of hemorrhage and there is a corresponding birth certificate within 42 days before the date of death, that woman likely died of a maternal death.

Again, the certified causes of death and the 42 day interval stem from the International Statistical Classification of Diseases and Related Health Problems

[9] http://www.fedora-commons.org/about
[10] http://www.irishgenealogy.ie/en/
[11] http://www.histpop.org/
[12] http://www.census.nationalarchives.ie/
[13] http://www.logainm.ie/
[14] http://dbpedia.org/About

(ICD). This contextual information is encoded in the ontology for data analytics and can change over time as the definition for maternal mortality is refined.

6 Conclusions and Future Work

In this paper we presented the IRL project, its aims and its conceptual architecture. From this point onwards; the ontology will be finalised and the digital archivist will start the process of archiving and digitally curating the vital records. The future steps are: i) to semi-automatically generate links between the same individuals in different records and links across datasets for transformation to RDF to provide additional contextual information; ii) develop methods to reconstruct virtual families; and iii) develop means to interrogate the data to answer the domain research questions. Given the practical nature of the last two points, we consider adopting a forward engineering method: design science [8]. The IRL platform aims to provide a solution for researchers – in this case historians – and the construction of such a solution needs appropriate methods for design, implementation and evaluation of a solution instead of falling in the trap as defining a practical problem as a knowledge problem.

Acknowledgements We thank the Registrar General of Ireland for permitting us to use this rich digital content contained in the vital records. 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. The Digital Repository of Ireland (formerly NAVR) gratefully acknowledges funding from the Irish HEA PRTLI programme.

References

1. Confidential Maternal Death Enquiry in Ireland, Report for Triennium 2009-2011. Cork: MDE (August 2012)
2. Bodenreider, O.: Disease ontology. In: Dubitzky, W., Wolkenhauer, O., Cho, K., Yokota, H. (eds.) Encyclopedia of Systems Biology, pp. 578–581. Springer New York (2013)
3. Brickley, D., Miller, L.: FOAF vocabulary specification 0.99 Namespace Document (January 2014), via `http://xmlns.com/foaf/spec/`
4. Fox, M.S., Gruninger, M.: Enterprise modeling. AI magazine 19(3), 109–121 (1998)
5. Geary, L.: Medicine and charity in Ireland, 1718-1851. University College Dublin Press (2004)
6. Grüninger, M., Fox, M.S.: The role of competency questions in enterprise engineering. In: BenchmarkingTheory and Practice, pp. 22–31. Springer (1995)
7. Jones, G.: Captain of all these men of death: the history of tuberculosis in nineteenth and twentieth century Ireland. No. 62, Rodopi (2001)
8. Wieringa, R.: Design science as nested problem solving. In: Vaishnavi, V.K., Purao, S. (eds.) DESRIST. ACM (2009)