Data Preservation and Interpretation Pipeline for Irish Civil Registration Records

Oya Beyan (a), PJ Mealy (a), Dolores Grant(b), Rebecca Grant (b), Natalie Harrower (b), Ciara Breathnach (c), Sandra Collins (b) and Stefan Decker (a)

(a) Insight @ NUIG, National University of Ireland Galway, Galway, Ireland
(b) Digital Repository of Ireland, Royal Irish Academy, Dublin, Ireland
(c) Department of History, University of Limerick, Limerick, Ireland

26 -30 October, 2015 @ META4eS’15 Rhodes, Greece
Motivation and Aim

• Civil registration records such as birth, death, and marriage registers contain a vast amount of implicit information about a society’s past.

• New knowledge can be generated 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 records for Dublin (1864-1913).
The **aim of the 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.

We have developed a **data preservation and interpretation pipeline** supported by a dedicated semantic architecture to

- explore and interpret historical data sets with semantic web technologies and Linked Data,
- serve diverse interest groups such as archivists, historians, journalists, public researchers and scholars,
- developed knowledge infrastructure that satisfies different and potentially conflicting perspectives and concerns of users
Data Set and Use Cases

- Data provided by the General Register Office under strict terms and conditions.

- Original data set contains of 6,009,781 births (from 1864 to 1912), 4,314,963 deaths (from 1864 and 1912) and 1,443,110 marriages (from 1845 to 1912).

- We demonstrated our concept with infant deaths as the use case.

- The pipeline will initially include 444 death register pages, which equates to 4090 death entries recorded in two Registrar Districts to the South of Dublin City from the years 1870 and 1890.
IRL: Irish Record Linkage, 1864 - 1913

**Data Preservation and Interpretation Pipeline**

**LAYER 1**
- Preserve the civil registers in their original form
- Capture the provenance of the archival record
- The Vital Records Ontology (VRO) was developed to annotate each register page
- Aim to preserve the authenticity in the original order and **without any interpretation**.

**LAYER 2**
- **Creating links** between the captured records and identifying the associations between them
- Includes annotations to other standards or ontologies such as the cause of deaths.
- The Historical Events Ontology (HEO) was developed to enrich the registers and interlink each archival entry to constitute families.

**LAYER 3**
- Exploring the linked records stored in the second layer from various points of interest
- Permits researchers to apply different definitions and **interpret the linked data**
- Use case specific ontologies can enable the historical data to withstand multi-factorial queries
Implementation of the IRL semantic pipeline

Data Acquisition

- Digital Archivist
- MySQL DB Instance

Preservation Layer

- D2RQ Mapping file using VRO
- Preserved file which illustrates the unchanged data represented in Vital Records Ontology

Interlinking & Enrichment Layer

- Data transformed to using HEO
- Domain information classification files.

Use Case Query Layer

- Java App (using JENA Api)
- Data with added relationships
- Jena Fuseki SPARQL Endpoint.
Data Acquisition

- The data consists of digitised birth, death and marriage register pages
- Access to the project data is restricted
- For the purpose of the project a MySQL database was created to curate a sample of the birth, death and marriage records from the Registrar Districts of Dublin South City 1 and Dublin South City 3.
- Digital archivists manually transcribed 444 death register page (4090 death records), 15 birth register page (150 birth records), and 28 marriage register page (81 marriage records) from 1870 and 1890.
- Death records have been focused on initially, as the historical research questions examine infant and maternal mortality.
Preservation Layer

- The aim is to provide a trustworthy platform for preserving the historical data by applying digital archival principles.
- Serves as a long-term digital preservation platform for digitised objects.
- Register pages are transcribed verbatim in the original form and represented in Linked Data based on the provenance and archival authenticity principles.
- D2RQ Mapping is applied to extract the data from the MySQL database into RDF using the VRO ontology.
- In the mappings, special care was taken to preserve the ability to trace information back to the source (the original records).
<table>
<thead>
<tr>
<th><strong>Register Page</strong></th>
<th><strong>Death Record</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="http://irl.dri.ie/register_page/D4746422" alt="Image" /></td>
<td><img src="http://irl.dri.ie/record/D4746422-69" alt="Image" /></td>
</tr>
<tr>
<td>records:RegisterPage ; rdfs:label &quot;D4746422&quot; ; records:county &quot;Dublin&quot; ; records:datePageCertified &quot;1890-01-02&quot;^^xsd:date ; records:datePageCertifiedAsTrueCopy &quot;1890-04-14&quot;^^xsd:date ; records:district &quot;South City Number 1&quot; ; records:districtOfSuperintendentRegistrar &quot;South Dublin&quot; ; records:forenameOfRegistrarOnPage &quot;752c4c2bdfb81a4e7511e7de&quot; ; records:forenameOfSuperintendentRegistrar &quot;2042101ac741bfe43f3672e67c&quot; ; records:pageNumber &quot;637&quot;^^xsd:int ; records:pageNumberOfManuscript &quot;1&quot;^^xsd:int ; records:quarter &quot;1&quot;^^xsd:int ; records:stampNumber &quot;4746422&quot; ; records:surnameOfRegistrarOnPage &quot;3f6d390dbfd5ab58b3109f6ba4&quot; ; records:surnameOfSuperintendentRegistrar &quot;76ecf24fdec371ebeb8e459c25d9f373&quot; ; records:union &quot;South Dublin&quot; ; records:volume &quot;2&quot;^^xsd:int ; records:withRecord <img src="http://irl.dri.ie/record/D4746422-67" alt="Image" />, <img src="http://irl.dri.ie/record/D4746422-61" alt="Image" />, <img src="http://irl.dri.ie/record/D4746422-62" alt="Image" />, <img src="http://irl.dri.ie/record/D4746422-68" alt="Image" />, <img src="http://irl.dri.ie/record/D4746422-65" alt="Image" />, <img src="http://irl.dri.ie/record/D4746422-70" alt="Image" />, <img src="http://irl.dri.ie/record/D4746422-63" alt="Image" />, <img src="http://irl.dri.ie/record/D4746422-66" alt="Image" />, <img src="http://irl.dri.ie/record/D4746422-64" alt="Image" /> ; records:yearRegistered &quot;1889&quot;^^xsd:int.</td>
<td>a records:Certificate , records:DeathRecord , records:Record ; rdfs:label &quot;Death of 5bd81ca81adf2879322e0fffd90b771c6db135761abfbeb3b2f79fc9ccba6 in 1889-12-29&quot; ; records:ageLastBirthday &quot;10 months&quot; ; records:causeOfDeath &quot;bronchitis&quot; ; records:condition &quot;bachelor&quot; ; records:dateOfDeath &quot;1889-12-29&quot;^^xsd:date ; records:dateOfRegistration &quot;1890-01-02&quot;^^xsd:date ; records:deathCertification &quot;Explicitly Certified&quot; ; records:durationOfIllness &quot;8 days&quot; ; records:forename &quot;5bd81ca81adar2879322e0fffd90b771c6db135761abfbeb3b2f79fc9ccba6&quot; ; records:forenameOfInformant &quot;341d3fba12d2c6c85e772e7be&quot; ; records:forenameOfRegistrar &quot;d005aa409ef3b5c6bf7cb6d8b41&quot; ; records:number &quot;69&quot;^^xsd:short ; records:placeOfDeath &quot;5 Lady Lane&quot; ; records:qualificationOfInformant &quot;present at death&quot; ; records:rankProfessionOrOccupation &quot;labourer's child&quot; ; records:residenceOfInformant &quot;ba5ae2bd62a87e6c9af264a3ef&quot; ; records:sex &quot;M&quot; ; records:surname &quot;c6db135761abfbeb3b2f79fc9ccba6&quot; ; records:surnameOfInformant &quot;c4a8bed037765363048185292&quot; ; records:surnameOfRegistrar &quot;7cb15d9c8b537a241da387619c&quot; ; records:titleOfRegistrar &quot;Registrar&quot;.</td>
</tr>
</tbody>
</table>
Interlinking and Enrichment Layer

• The aim is to facilitate the exploitation of historical data for various purposes by enabling efficient queries.
• VRO designed for preserving digitised objects as they are. Therefore, 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.
• HEO defines objects for the actors of 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.
IRL: Irish Record Linkage, 1864 - 1913
Interlinking and Enrichment Layer

• The interlinking and enrichment layer was implemented with a Java application using JENA API.

• VRO based linked records were processed and converted to a new data model based on the HEO and transform to another triple store for exploring the data through generations.

• 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.

• The original records are enriched by adding derived features. E.g. the addition of an ageAtDeathInMinutes field.
IRL: Irish Record Linkage, 1864 - 1913

<http://irl.dri.ie/record/D4746422-69/person>

  a heo:Person ;
  rdfs:seeAlso "http://irl.dri.ie/record/D4746422-69" ;
  heo:AgeAtLastBirthday "10 months" ;
  heo:AgeAtLastBirthdayInMins "439200" ;
  heo:CondAtDeath "bachelor" ;
  heo:dateOfDeath "1889-12-29" ;
  heo:forename "5bd81ca81adf2879322e0ff9d90b771" ;
  heo:hasAtDeath <http://irl.dri.ie/record/D4746422-69/rank> ;
  heo:hasCauseOfDeath [ heo:classifiedAs "http://purl.org/net/irish-record-linkage/historicalEvents.owl#Bronchitis" ;
    heo:durationOfIllness "8 days" ;
    heo:originalText "bronchitis"
  ] ;
  heo:hasRecordFor <http://irl.dri.ie/record/D4746422-69/deathEvent> ;
  heo:surname "c6db135761abfeb3b2f79fcb9ccba6" .

Dead Person
<http://irl.dri.ie/record/D4746422-69/deathEvent>
  a heo:DeathEvent ;
  heo:InformantQualification "present at death" ;
  heo:dateOfRegistration "1890-01-02"^^<http://www.w3.org/2001/XMLSchema#date> ;
  heo:deathCertification "Explicitly Certified" ;
  heo:eventOf <http://irl.dri.ie/record/D4746422-69/person> ;
  heo:placeOfDeath "5 Lady Lane" ;

<http://irl.dri.ie/record/D4746422-69/informant>
  a heo:Person ;
  heo:forename "341d3f1faa12b2ccbb5e73da72e7be" ;
  heo:relationsWithSubject "NO RECORDED relationshipInformant" ;
  heo:surname "c4a648becd03776536304811c85292" .

<http://irl.dri.ie/record/D4746422-69/registrar>
  a heo:Person ;
  heo:forename "d0c05aa409ef3b5c6bf7cb601d8b41" ;
  heo:surname "7cb15d9c8b537a241a1dca3847619c" .
Interlinking and Enrichment Layer

- The data set is enriched with standard terminologies and ontologies. E.g. cause of death is mapped to various coding systems.
- Reflecting significant advances in medical science, medical coding systems underwent a similar evolution in the period under review.
- Three available coding systems are selected namely,
  - the International List of Causes of Death, Revision 1 (1900) (ILCD1),
  - the International List of Causes of Death, Revision 2 (1909) (ILCD2),
  - the International Classification of Causes of Sickness and Death (ICSD)
Use Case Query Layer

• Aim 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?”

• JENA Fuseki SPARQL endpoint serves to address the use cases and return the query responses.

• Historic definitions vary for maternal and infant mortality. Infant mortality is currently defined as a death of a child before reaching the age of one.

• Deaths in the first 24 hours and in the following 27 days have specific significance from the historians’ perspective.
Use Case Query Layer

- 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.

```sql
prefix heo: <http://purl.org/net/irish-record-linkage/historicalEvents.owl#>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
select ?s ?DateOfDeath ?AgeInWords
Where {
  ?s a heo:Person.
  ?s heo:AgeAtLastBirthdayInMins ?ageInMins.
  ?s heo:AgeAtLastBirthday ?AgeInWords .
  ?s heo:dateOfDeath ?DateOfDeath
  FILTER(xsd:integer(?ageInMins) <= 1440) #1440 is 24 hours
  FILTER(xsd:integer(?ageInMins) > 0) #ignore records with no recorded age }
```
Reflexions and Future Work

- Semantic technologies and Linked Data promises many advantage for capturing, exploring and interpreting historical data sets.
- The domain ontologies may provide means for separating varying concerns, preserving authenticity and maintaining the provenance of the records.
- In this work we
  - explored the application principles of the semantic web technologies to create a data preservation and knowledge query pipeline by utilising the linked data,
  - developed supporting domain ontologies, namely VRO and HEO.
IRL: Irish Record Linkage, 1864 - 1913

Reflexions and Future Work

• Explore methods for more flexible and dynamic use case generation and user friendly applications for end users.
• Investigate to what extent parts of the knowledge bases can be made available to the public without revealing the details of individuals.
• Develop techniques to improve scalability and performance of the developed technologies.
• On Ontology Engineering. The notes and anomalies kept by the digital archivists can be used to create a controlled vocabulary for anomalies. This leads, in turn, to extra predicates for information retrieval.
Questions?

More information
• Twitter: @IRL_Project
• Project website http://irishrecordlinkage.wordpress.com/