Noname manuscript No.
(will be inserted by the editor)

A Semantic Architecture for Preserving and Interpreting the Information Contained in Irish Historical Vital Records

Christophe Debruyne · Oya Deniz Beyan · Rebecca Grant · Sandra Collins · Stefan Decker · Natalie Harrower

Received: date / Accepted: date

Abstract Irish Record Linkage 1864-1913 is a multi-disciplinary project that started in 2014 aiming to create a platform for analyzing events captured in historical birth, marriage and death records by applying semantic technologies for annotating, storing and inferring information from the data contained in those records. This enables researchers to, among other things, investigate to what extent maternal and infant mortality rates were underreported. We report on the semantic architecture, provide motivation for the adoption of RDF and Linked Data principles, and elaborate on the ontology construction process that was influenced by both the requirements of the digital archivists and historians. Concerns of digital archivists include the preservation of the archival record and following best practices in preservation, cataloguing and data protection. The historians in this project wish to discover certain patterns in those vital records. An important aspect of the semantic architecture is the clear separation of concerns that reflects those distinct requirements – the transcription and archival authenticity of the register pages and the interpretation of the transcribed data – that led to the creation of two distinct ontologies and knowledge bases. The advantage of this clear separation is the transcription of register pages resulted in a reusable dataset fit for other research purposes. These transcriptions were enriched with metadata according to best practices in archiving for ingestion in suitable longterm digital preservation platforms.

Keywords Historical Vital Records · Cultural Heritage · Linked Data · Ontology Engineering · Preservation

We are grateful to the Registrar General of Ireland for permitting us to use the 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. The Digital Repository of Ireland (formerly NAVR) gratefully acknowledges funding from the Irish HEA PRTLI programme. We also would like to thank Prof. Declan O'Sullivan from the ADAPT Centre at Trinity College Dublin and the anonymous reviewers for their valuable feedback. Christophe Debruyne is currently supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of the ADAPT Centre for Digital Technologies Platform Research at Trinity College Dublin.

C. Debruyne
ADAPT Centre, Trinity College Dublin, Dublin 2, Ireland
WISE Lab, Vrije Universiteit Brussel, 1050 Brussels, Belgium
Tel.: +353 1 896 4992
E-mail: debruync@scss.tcd.ie

O. D. Beyan
RWTH Aachen University, 52062 Aachen, Germany
E-mail: beyan@dbis.rwth-aachen.de

R. Grant
Royal Irish Academy, Dublin 2, Ireland
E-mail: r.grant@ria.ie

S. Collins
National Library of Ireland, Dublin 2, Ireland
E-mail: scollins@nli.ie

S. Decker
RWTH Aachen University, 52062 Aachen, Germany
Fraunhofer FIT, Schloss Birlinghoven, 53754 Sankt Augustin, Germany
E-mail: decker@dbis.rwth-aachen.de

N. Harrower
Royal Irish Academy, Dublin 2, Ireland
E-mail: n.harrower@ria.ie

1 Introduction

The multi-disciplinary Irish Record Linkage (IRL) 1864-1913 project[1] aims to provide a comprehensive map of

[1] https://irishrecordlinkage.wordpress.com/

infant and maternal mortality for Dublin from 1864 to 1913. The project aims to reconstruct family units and create longitudinal histories by linking records of birth, marriage and death (vital registration data) provided by the General Register Office [7].

In order to create and analyze *meaningful* links across and between the different entities captures in those vital records, the project created a knowledge platform which adopted Semantic and Linked Data technologies. To enhance the research potential of the dataset that is being developed, best practices in digital archiving and digital preservation have been taken into account next to fulfilling the information needs of the historians.

In this paper, we report on the semantic architecture and ontology creation; the creation of a knowledge base containing historical birth-, marriage- and death records translated into RDF; the creation of a Linked Data [10] platform to aid historians in analyzing those events; and how we preserve the information captured in those records in suitable longterm digital preservation platforms.

The project involves the expertise of three disciplines [7]: historians, digital archivists and knowledge engineers. With the help of knowledge engineers creating the ontologies and setting up the platform and the digital archivists who curate, ingest and maintain the RDF, the historians will be able to analyze 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. To aid the historians in their data analysis, the knowledge engineers also contribute in linking people across records and the contextualization of the information with other datasets. Both knowledge engineers and digital archivists also collaborated in proposing an information-processing pipeline to enrich the data using suitable metadata formats to facilitate exploration and discovery in suitable longterm digital preservation platforms.

The development of our Linked Data platform was driven by several questions. First, rather than creating a dataset fit for the research questions posed by the historians, we wanted to investigate how digital archivists can create a dataset that can be reused for various purposes by capturing and transcribing the contents of these records which can then be reused for testing specific hypotheses. Secondly, to enhance the data's potential, we also aimed to investigate how the dataset created by the digital archivists can be digitally preserved in a longterm preservation platform.

The main contributions of this paper – which extends work reported in [18] – are an elaboration of the notion of separating transcriptions of artifacts and the interpretation thereof, which we call separation of concerns; the transformation from the RDF transcriptions into RDF graphs that aid the historians in answering their research questions; and how the platform allows one to distill RDF transcriptions of the digitized objects that can be ingested in a suitable longterm digital preservation platform. For the latter, metadata records are distilled from the RDF transcriptions that are used to archive, explore and discover the records on said platform.

The remainder of this paper is organized as follows: we first present the General Register Office and describe the register pages in Section 2, we then proceed with an overall description of the IRL Semantic Platform in Section 3. Section 3 is followed by descriptions of different aspects of the platform: the transcription of register pages by digital archivists in Section 4, the creation of RDF representations of those register pages in Section 5, the interpretation of that RDF by enriching and contextualizing the information in Section 6, and the ingestion of the generated RDF – together with the digitized versions of those register pages and a metadata record for archiving purposes – in a longterm preservation platform in Section 7. Prior to concluding this paper in Section 10, we provide a discussion and relate our contributions with respect to the state of the art in Sections 8 and 9 respectively.

2 General Records Office

In Ireland, the General Register Office (GRO) is Ireland's civil registry responsible for recording information on births, deaths and marriages. In this project, the Registrar General of Ireland generously offered us records 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)

under strict terms and conditions. It became compulsory to report and register births, deaths and marriages in 1864, but *non-Catholic* marriages were already being registered from 1845 onwards.[2] This explains the broader timespan for marriage records. *Records* of these events were captured on *register pages* (up to 10 per page for births and deaths, and up to 4 for marriages) divided by district and sent to the GRO where volumes were then created and an *index* compiled. The information was provided to us as a database dump of the

[2] `http://www.irish-genealogy-toolkit.com/Irish-marriage-records.html`

GRO's database with digitized versions of the register pages and indexes.[3]

The information system built by the GRO allows to search for vital records of persons based on a person's name, geographical area (to the level of district) and year; one of their core services to the public. Not only has the GRO spent resources in the construction of such a service, an enormous amount of effort also went into the digitization of register pages and indexes as accurately as the recording of a subset of the information in a relational database. A rational decision was made to only enter in the database the information necessary to efficiently find records.

While the system developed by the GRO works perfectly for finding historical records, information that is key in answering the IRL historians' questions were not captured by the database (such as the places of death, names of the informant, etc.). The development of a platform fit for the historians' needs would not only require the addition of missing fields. The register pages contain information beyond their form-structure (such as additions, anomalies, crossed out information, etc.) that is worth capturing in some way. There is furthermore a lot of implicit information in those register pages which can be made explicit in a meaningful way (e.g., the different relationships between people). To model and capture this information in preparation for the Linked Data platform to be developed, one should call upon the expertise of knowledge engineers and digital archivists – skills that are typically not present in an organization – who will work in tandem. Digital archivists are trained in processing, transcribing and curating the information (which includes making informed decisions about choosing, reusing, building controlled vocabularies, among other things). Knowledge engineers are skilled in capturing and organizing domain knowledge to solve complex problems. Knowledge engineers with Linked Data expertise are furthermore capable of building knowledge-based systems that are linked with other datasets to enrich and contextualized the information.

The vital records and the goals of the IRL project lead to various challenges that need to be taken into account and those challenges reside at different levels: data protection, data transcription, historical evolution (medical knowledge, geographical, etc.) and, of course, the method for answering the historians' research questions. We will highlight some of the pertinent challenges below that will influence the design of the semantic architecture and the transcription workflow.

Data security and protection in terms of transfer, storage and use by authorized parties were covered by the data sharing agreement with the GRO. The goal of the IRL project is to build a platform that allows one to analyze the data captured in those records and not to replace the service already built by the GRO, although the new platform would support the queries typically executed by GRO as well. As per our data sharing agreement, the dataset in its entirety (that means the data and the digitized objects) should only be available to the members of the project team. With the help of the digital archivists, who are familiar with data protection legislation and best practices, we furthermore identified which guidelines to follow.

Records, knowledge and interpretation. A second challenge is the varying levels of detail in the records (seen in, for instance, the causes of death) and the variances in how subject names and places were recorded (initials, short hands, name of a building versus street name, etc.) [7]. These variances might imply something, which we are currently unaware of. Therefore, we should ensure that the transcription of the register pages transcribes exactly what was written down. In other words, the manipulation of the information should be kept to a minimum. This leads to another, yet related challenge, *clearly separate two concerns* [7]: the exact transcription of what has been captured on the register pages as to have an authentic virtual account of historic events; and the interpretation, possibly with background knowledge, of certain aspects based on these interpretations. Examples of how interpretation can differ are the evolution of Ireland's geography (place names changing and streets disappearing, merging and even reappearing), evolution in knowledge (e.g., new insights in medicine) and even the adoptions of different theories (e.g., different classifications of social status). This separation of concerns was also elaborated a survey paper on the application of Semantic Web technologies for historical research published in 2015 [27] where they argued that data transformations should adhere to two constraints: i) keeping the original source data intact, ii) and storing changes to data in separate artefacts and keeping track of the changes made.

Provenance and archival authenticity. Archival theory is based on two key principles, *respect de fonds* (original order) and *archival provenance*. The first is the principle which guides archivists when exerting intellectual control over a collection, and ensures that

[3] The terms and conditions of our data sharing agreement do not permit us to make public any data that would identify any individual [7]. One can 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.

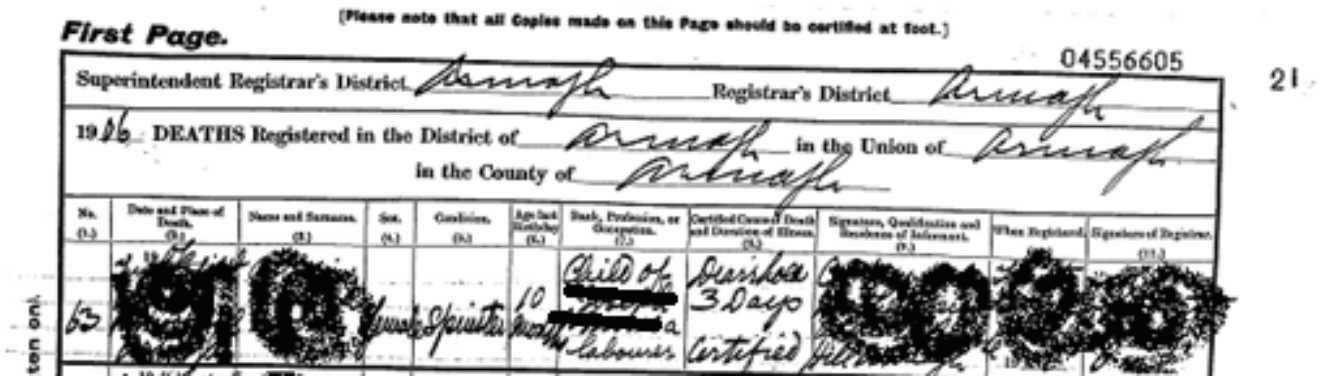


Fig. 1 Part of a register page containing death records (redacted as per our data sharing agreement). Copyright held by the General Register Office and reproduced with permission.

the archival record is always described in relation to the context in which it is created as far as possible (for example a letter should only be described in terms of a set of correspondence where it is available). 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. The principle of *respect de fonds* is linked closely to provenance, which forms the foundation of archival description. Provenance refers to how the archival record relates to its creator, and can only be maintained through the appropriate description of an archival record. These principles are important in the digital sphere, and describing and authenticating records in this way gives meaning through the provision of context.

Other data challenges include the conversion to appropriate data formats as well as cataloguing of the digitized objects so as to ensure compliance with digital preservation best practices. These challenges, however, fall outside of the scope of this paper; work on the ingestion of the digitized objects in a suitable longterm digital preservation platform will be disseminated elsewhere.

3 IRL Semantic Architecture

The semantic architecture of the IRL platform is set up to cope with the requirements defined by the data challenges described in the previous section and the research questions the historians aim to address. Fig. 2 depicts graphically our architecture in which the two aforementioned concerns – exact transcription on the left vs. interpretation on the right – are strictly separated.

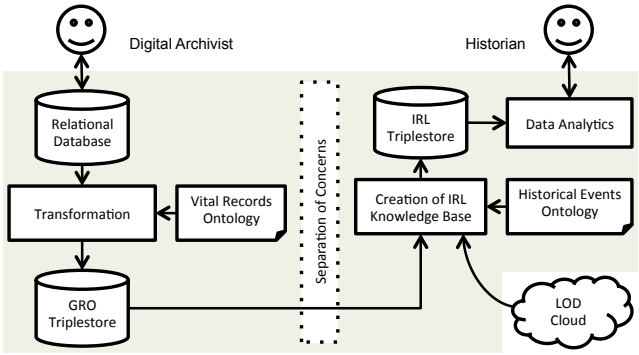


Fig. 2 The conceptual architecture of the IRL Linked Data platform. Transcription of register pages and the interpretation of the data are strictly separated. Note that the Separation of Concerns presented in this diagram is not a component of the system, but a principle adhered to in the IRL project.

Digital archivists transcribe register pages by populating a relational database[4] with an interface[5]. The schema of that database closely follows the structure of the register pages. The advantage of a relational database is that the table definitions allow for certain constraints to be satisfied before data can be entered. This already allows for some quality assurance in terms of valid data entry during transcription.

We will first describe the motivation for the adoption of RDF and semantic technologies and discuss some aspects of each concern. Details on the ontologies developed for this platform will be discussed in subsequent sections and build further upon the work reported in [7].

RDF and Linked Data principles were adopted for various reasons. RDF allows us to use a simple data model that facilitates the integration of internal and external data by creating links. Using RDF, the management of knowledge is scalable, and data access –

[4] A MySQL database (`https://www.mysql.com/`).
[5] With phpMyAdmin (`https://www.phpmyadmin.net/`).

for analysis, among other things – is pushed closer to the user and application level by adopting the Linked Data principles (e.g., content negotiation) and the W3C SPARQL recommendation. As [31] noted, the adoption of Semantic Web technologies allows one to easily build applications for different stakeholders on top of RDF stores via SPARQL.

By reusing the already existing HTTP infrastructure on which Linked Data is built, datasets that are behind firewalls can still link to other datasets in the Linked Data cloud. This allowed us to take a conservative approach by setting up our services behind a firewall and create (and exploit) outbound links; we thus benefit from all the Semantic Web technologies and the Linked Data cloud has to offer without violating our data sharing agreement. Datasets relevant for this project that provide additional context include DBpedia [4] and Linked Logainm [25]. The latter is a Linked Data version of the authoritative bilingual database of Irish place names `logainm.ie` and provides links to places in DBpedia and `geonames.org`.

OWL 2 was adopted for the creation of the two ontologies allowing us to infer implicit information and rule languages were adopted to encode domain expert knowledge (historical, medical, etc.) to infer additional information that falls outside the capabilities of OWL. Other advantages of adopting OWL 2 were reported in [16]: the use of inverse properties allows one to explore resources from either side of a relationship and the ontologies can be easily extended with extra information to suit other types of applications.

There are four principles that Linked Data datasets should adhere to [6]: 1) use URIs as names for things; 2) use HTTP URIs so that people can look up those names; 3) provide information with standards (such as RDF) when URIs are looked up; and 4) include links to other URIs. Principles 1 to 3 are adhered to by both triplestores. The GRO triplestore provides links to other URIs within the same dataset to avoid interpretation and contextualization. The IRL triplestore links to external datasets to provide that contextualization. Since the datasets are behind a firewall, inbound links are not possible. Outbound links can be followed to discover more information. The authors are aware that the firewall can pose problems if one wishes to execute federated queries (across different datasets), but this has not yet been encountered within the context of this project.

For the platform, we adopted Jena TDB as triplestores and Jena Fuseki to provide the SPARQL endpoints.[6] Pubby is used to create a simple Linked Data frontend via those endpoints.[7] Details on the technologies adopted for the generation of RDF triples from the relational database and the transformation of triples for the interpretation of the data will be provided in the next sections.

Finally, the historians who wish to answer certain research questions will use a set of tools that will aid them in analyzing the data contained (in an implicit or explicit manner) in the register pages. They have access to the second triplestore with data interpreted in certain ways. Questions they wish to see answered are formulated either in terms of rules (e.g., SWRL[8]) or as SPARQL queries.

Boonstra *et al.* presented the historical information lifecycle in which the phases are [11]: creation, enrichment, editing, retrieval, analysis and presentation. The end of the presentation phase can trigger a new creation phase, closing the loop and therefore also providing an iterative lifecycle. The names of each phase are self-explanatory and – according to the authors – do not have to be rigorously followed. Enrichment is concerned with enriching the created data with metadata to facilitate information retrieval and discovery (e.g., with Dublin Core). Editing is concerned with both the actual encoding (annotation) and algorithmic transformations. We note that in our platform, both editing and enrichment happens at two "levels" since we explicitly separate two concerns. Digital archivists transcribe (or "edit") the data in a relational database including provenance information (considered "enrichment"). These transcriptions are subsequently "edited" a second time for a different purpose with transformations using a series of SPARQL CONSTRUCT queries and rules and creating links with other knowledge sources ("enrichment").

4 Transcription of the Register Pages

We reiterate that the existing system the GRO has built took into account the attributes necessary to find records about individuals, thereby leaving out all fields on the register pages that were not relevant for this task. The digital archivists thus have the meticulous and laborious task of transcribing all the data that was captured on register pages, which is not merely transcribing those records, but also involves undertaking research and controlling the quality of what has been transcribed. The adoption of Optical Character Recognition (OCR) was not possible as a very high level of precision in the transcription process was necessary.

[6] `http://jena.apache.org/`

[7] `http://wifo5-03.informatik.uni-mannheim.de/pubby/`

[8] `https://www.w3.org/Submission/SWRL/`

In order to cope with the tension field of transcribing exactly what has been written down and the normalization of the data in some of these fields, a relational database has been set up that can capture in greater detail what can be observed on a register page. On death register pages, for instance, one can find a field "Certified Cause of Death and Duration of Illness". We observed variances in detail, which depended for instance on the registrar or on the informant (practitioner vs. relative). That field was sometimes used to indicate that the cause of death was uncertified. The database thus provided an additional field to indicate whether a death was explicitly certified, explicitly uncertified or neither. The duration of illness can be unknown or not applicable, e.g., in the case of drowning. The field can thus be NULL in case no information was provided.

Notes for each record and register page can be kept to capture anomalies or peculiarities such as signatures with a cross or crossed out information. As the project continues and the digital archivists transcribe register pages, these notes could be used as input for the creation of a controlled vocabulary for anomalies in register pages (see future work). The database schema was developed in such a way that the data entered adheres to certain integrity constraints, thus effectively preventing certain errors.

Many tools have emerged [33] to generate RDF from these databases either via direct mappings – based on the ideas outlined in [5] with the RDF reflecting the database's structure and labels – or via mappings where tables, views and queries are related to concepts and relations in ontologies to create a knowledge base. We deemed the second approach more viable for several reasons. First, it does not create a dependency on the database and its structure. The structure and even the database can change while the ontology remains stable and the only things to update are the mappings. Secondly, the use of ontologies allows one to map several non-RDF datasources with the same ontology.

Our relational database is annotated with the Vital Records Ontology, presented in the next section, using D2RQ [9] and the generated triples are stored in a records triplestore. Though both approaches ultimately led to two W3C Recommendations, a direct mapping of relational data to RDF [3] and R2RML [17], we have chosen to adopt D2RQ as it came with a built-in Linked Data front end and SPARQL endpoint facilitating the development, testing and exploration of the generated triples.

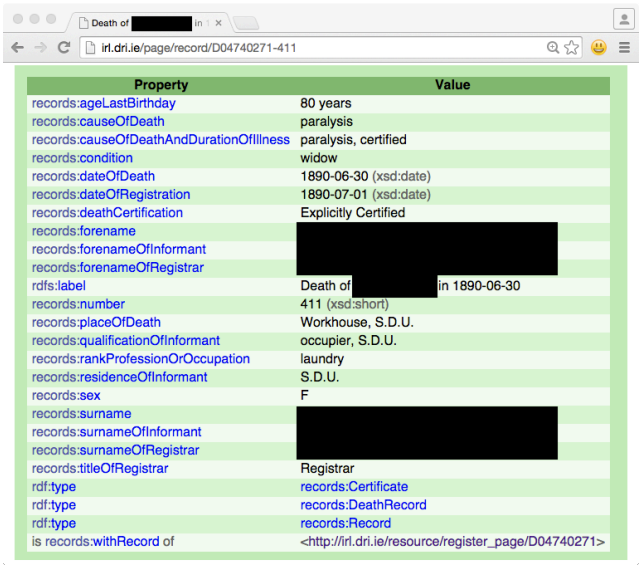
5 Vital Records Ontology (VRO)

Births, deaths and marriages were captured per district (within a union, within a county) as single records on register pages. These pages can contain up to 10 records after which such a page is signed off by the registrar and sent to the superintendent registrar for inspection and validation. To create a first version of the Vital Records Ontology (VRO)[9], we "lifted" the information one could see on one such register page to an ontology.

To minimize 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. For example, instead of creating a concept `Person` that can have a `forename` and `surname`, we choose to relate the concept of a `Record` to these attributes. For the VRO, we thus defined a few concepts. A `RegisterPage` and a `Record` for representing the different types of records were declared. Each record must belong to a register page and each register page can have zero (which implies a blank pages) or more records. We make a distinction between a `Certificate` and a `MarriageRecord`, both of them being disjoint subclasses of the concept `Record`. The first has as a subject only one person and the latter two. The two concepts are disjoint, which makes that no instance of a certificate can be an instance of a marriage record and vice versa. Finally, we created two disjoint subclasses of the concept `Certificate`: `BirthRecord` and `DeathRecord`. The only object property, a relation between two concepts, we needed was to relate records to register pages. All other properties are datatype properties. Datatype properties are related to the greatest common denominator. For instance, all records are signed off by a registrar on a certain date. The date of registration as well as information on the registrar are therefore related to the concept of `Record` so that all subtypes of this class inherit this property.

One of the challenges is to capture the domain as well as possible, yet maintain a valid OWL 2 ontology. This is to ensure that we can support complete reasoning over the knowledgebase for subsequent data analysis purposes. As explained by Motik and Horrocks in [28], it is difficult to reason about date and time intervals, and therefore only specific points in time (captured by both `xsd:dateTime` and `xsd:dateTimeStamp`) were "amenable for implementation" and those "can be handled by techniques similar to the ones for numbers." Together with the digital archivist, we choose not to capture dates mentioned in records as instances of the `xsd:dateTime` datatype as we do not know the

[9] Available via `http://purl.org/net/irish-record-linkage/records`.



Property	Value
records:ageLastBirthday	80 years
records:causeOfDeath	paralysis
records:causeOfDeathAndDurationOfIllness	paralysis, certified
records:condition	widow
records:dateOfDeath	1890-06-30 (xsd:date)
records:dateOfRegistration	1890-07-01 (xsd:date)
records:deathCertification	Explicitly Certified
records:forename	[redacted]
records:forenameOfInformant	
records:forenameOfRegistrar	
rdfs:label	Death of [redacted] in 1890-06-30
records:number	411 (xsd:short)
records:placeOfDeath	Workhouse, S.D.U.
records:qualificationOfInformant	occupier, S.D.U.
records:rankProfessionOrOccupation	laundry
records:residenceOfInformant	S.D.U.
records:sex	F
records:surname	[redacted]
records:surnameOfInformant	
records:surnameOfRegistrar	
records:titleOfRegistrar	Registrar
rdf:type	records:Certificate
rdf:type	records:DeathRecord
rdf:type	records:Record
is records:withRecord of	<http://irl.dri.ie/resource/register_page/D04740271>

Fig. 3 Example of the triples from a death record in a register page.

exact times and we felt that encoding "default" times would not be in keeping with archival principles. We thus chose to declare the range of these properties as being `rdfs:Literal`, but provided transcription guidelines in which the use of `xsd:date` was to be highly encouraged.

One key requirement for Linked Data platforms in general is adequate identifiers. For our records knowledge base, we need to identify instances of records and register pages. Each register page and record is identified by a URI under a new subdomain[10]. Register pages are identified by a unique, physically stamped number provided by the GRO while digitizing. We use this stamp number for the creation of URIs identifying register pages. Individual records are identified by the combination of the stamp and entry-number. Fig. 3 depicts the triples from a death record on a register page of a woman who died of paralysis in the year 1890.

6 Interpretation of the Register Pages and Records

We already described the importance of separating the information captured in the register pages and the interpretation thereof. The ontology that needs to support that kind of interpretation of the GRO data is more challenging given that the historians wishing to analyze the content are not necessarily familiar with ontology engineering and the knowledge base needs to support their activities, we adopted – reported in [7] – the approach proposed by Grüninger and Fox of having the stakeholders formulating *competency questions* [24]. The ontology must contain a necessary and sufficient set of axioms to represent and solve these questions [24]. These competency questions are not used to generate an ontology, but rather to evaluate it [21]. 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 included (paraphrased from [7]): "How many women died within n days after childbirth due to complications related to labor [...]?" and "What is the average sibship interval where the first child did not survive under various socio-economic conditions?" Those questions can be broken down in smaller competency questions such as: "Which infants died within the first 24 hours of their life?" and "What was the cause of death of a person?"

The questions were analyzed to identify the concepts and relations for the ontology, which were validated by the stakeholders. Graphical representations of the developed ontologies were used during discussions, e.g., as shown in Fig. 4. The VRO serves to reflect the historical records. Although it contains information about *events*, *people*, *places*, etc., the VRO does not capture these as *distinct entities*. However, to reconstitute families and analyze, we need distinct representations of events and persons involved. Therefore we developed the *Historical Events Ontology* (HEO) on top of the VRO as to provide a base ontology for answering the competency questions. The choice was made not to declare these concepts in the VRO as they fulfill the requirement of one particular set of tasks. This strict **separation of concerns** would allow for a **greater reuse of the historical records** for different kinds of analyses.

6.1 Historical Events Ontology

The Historical Events Ontology (HEO) was developed to reconstruct families with a life course perspective and enable the effective querying of competency questions for historians. Within this contextual frame two major concepts are identified: `Person` and `Event`. The first represents anyone participating in the actualization or recording of the latter. When we examined the registry pages, we observed that these pages provide a rich source for describing various persona roles. They can be subject or object of an event, they can be witness or recorder of an event, or they could be the one who played direct or indirect role in the occurrence of

[10] `http://irl.dri.ie/`

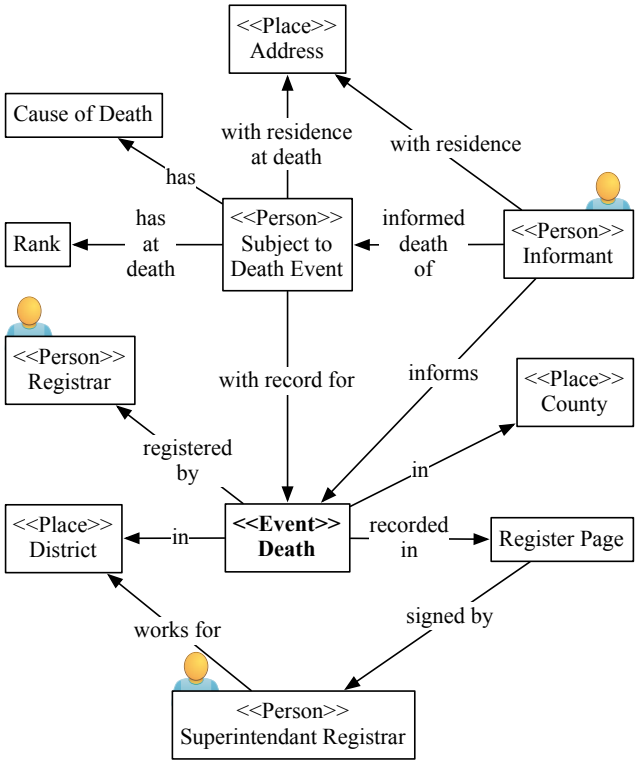


Fig. 4 Concepts and relations in the Historical Events Ontology for deaths.

an event such as father. To represent these rich participation relations a set of object properties were created between `Person` and `Event` including: `hasInformant`, `registeredBy`, and `hasRecordFor`. The concept of an event is furthermore specialized into the concepts of a `BirthEvent`, a `DeathEvent` and a `MarriageEvent`. Each specialization corresponds to exactly one of the certificates/records concepts in the VRO ontology and cannot be expended without adding any new historical data sources. Fig. 5 demonstrates how people, events, records and register pages are related with the HEO ontology.

Another important feature of the HEO is that it enables the description of the nature of relations between persons to reconstruct the family relations with a temporal dimension. Therefore we were required to define object properties from `Person` to `Person` for representing kinship relations captured by registry pages. We looked at existing ontologies for reuse and integration as well as the creation of missing concepts and relations for the creation of the HEO. To describe people and their relations, we take into account FOAF[11] and the Persona Vocabulary[12]. Both are used to describe people, their activities and their relations to other people and objects. However in FOAF vocabulary, many kinds of relationships between people are deliberately simplified as `knows`. The latter has more relations such as `hasChild`. In the HEO ontology we reused them to describe relations.

Concepts in the HEO include: `Person` for those involved during the event or registration; `Event` to capture the recorded births, deaths and marriages; `Place` for locations related to events or people; `CauseOfDeath` to facilitate reasoning over and classifying causes of death; `Rank` for capturing the rank and occupation of involved persons; and `RegisterPage` to assure provenance. In a first instance, the data from the first triplestore is transformed to populate concepts and relations in the HEO by a series of SPARQL CONSTRUCT queries and SWRL rules. For instance, the following query allows us to create instances of `foaf:Person` from death records (prefixes omitted):

```
CONSTRUCT {
 ?new a foaf:Person;
      rdfs:seeAlso ?r;
      foaf:firstName ?f;
      foaf:familyName ?s.
}  WHERE {
 ?r a rec:DeathRecord;
    rec:forename ?f;
    rec:surname ?s.
 BIND (URI(CONCAT(STR(?r),"/person")) AS ?new).
}
```

Transforming graphs from the first knowledge base into the second leads to the creation of many persons. Matching techniques are adopted to identify the same persons across different vital records to assert facts with the `owl:sameAs` predicate. This is an important as some names are very common and women adopted the name of their husband after marriage. Other fields (place, time) need to be taken into account to properly identify the same persons across records. When transforming graphs from the first knowledge base into graphs for the second, many instances of person are created. Another goal of the IRL platform is to add contextual information from other datasets [7]. We adopted Linked Logainm [25] for information on Irish place names and links with DBpedia resources.

6.2 Enrichment and Interpretation

As the project aims to reconstitute families and health histories of people, we also included concepts related to time (events), relations, and reused available domain disease ontologies. The construction of the HEO also included formalizing information found in classification

[11] Friend-of-a-Friend: `http://xmlns.com/foaf/spec/`
[12] `http://wiki.eclipse.org/Persona_vocabulary`

```
<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 "5bd81ca81adf2879322e0ffd90b771" ;
  heo:surname "c6db135761abfeb3b2f79fcb9ccba6'' ;
  heo:hasAtDeath <http://irl.dri.ie/record/D4746422-69/rank> ;
  heo:hasRecordFor <http://irl.dri.ie/record/D4746422-69/deathEvent> ;
  heo:hasCauseOfDeath [
    heo:classifiedAs "http://purl.org/net/irish-record-linkage/historicalEvents.owl#Bronchitis" ;
    heo:durationOfIllness "8 days" ;
    heo:originalText "bronchitis"
  ]
  .
```

Fig. 5 An example of a deceased person, related to an instance of a death records via the `rdfs:seeAlso` predicate. Again, forenames and surnames have been obfuscated as per our data sharing agreement.

systems such as the International Statistical Classification of Diseases and Related Health Problems.[13]

Another type 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 [1]. 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)[14], the International List of Causes of Death, Revision 2 (1909) (ILCD2)[15], and the International Classification of Causes of Sickness and Death (ICSD)[16]. The distinct cause of death is selected from the triple store, manually reviewed by the domain experts, and mapped to the available codes in ILCD1, ILCD2, and ICSD.[17]

In HEO, we created `CauseOfDeath` and identified subcategories for each of them. Each subcategory is

[13] `http://apps.who.int/classifications/icd10/browse/2010/en`

[14] Int. List of Causes of Death, Rev.1 (1900). `http://www.wolfbane.com/icd/icd1h.html`

[15] Int. List of Causes of Death, Rev.2 (1909). `http://www.wolfbane.com/icd/icd2h.html`

[16] Department of Commerce and Labor, Bureau of Census. International Classification of Causes of Sickness and Death. Washington Government of Printing Office (1910)

[17] We used the classification systems that existed in the studied historical period rather than applying today's most current classification systems because classification systems reflect a different understanding of disease than those in the 19th century. Diseases may be classified by etiology (cause), pathogenesis (mechanism by which the disease is caused), or by symptom(s). Nosology is a branch of medicine deals with classification of disease. The historical evolution of classification systems, such as ICD or ICSD, is closely related with historical and intellectual conditions of the area. The Early disease classification used by physicians was largely based philosophically on humoral theories of disease, with occasional suggestions that malign outside influences might cause illness or death. The first version of ICD included the principle of classifying diseases by etiology. In later years, the focus first shifted to symptoms and then to mechanism of diseases. For example in the historical records we observed "Teething" as cause of death. International List of Causes of Death, Revision 1 provides a classification category for this such as "82 Teething" for infants. The latest version of same classification (ICD10 or ICD11) does not have such a category as a disease or cause of death. A second reason for adopting historical classification systems is the number of categories that have expanded dramatically to reflect the new insights for understanding cause, mechanism and symptoms of diseases as medical knowledge advanced. The first version of International List of Causes of Death, Revision 1 (1900) had 191 items, whereas current one has more than 14,400 different codes. Mapping the historical disease classification to current ones would require the examination of historical definitions of each category and map each of them to current possible understanding of diseases. In such a mapping, a historian can explore how medical knowledge and social conditions effects the formation of nosologies, but it would not have served our purpose of classifying historical cause of diseases in 19th century.

annotated with the relevant ILCD1, ILCD2 and ICSD codes. As shown in Fig. 5, 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 in HEO. 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. 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.

We also derive information to facilitate querying. One example is `AgeAtLastBirthdayInMins`, for capturing the age at death (captured in register pages in terms of months, minutes, years, etc.) in minutes such that a homogenous representation is available for all events.

6.3 Retrieving Answers to the Competency Questions

JENA Fuseki SPARQL endpoint serves to address the use cases and return the query responses. The ultimate aim of the semantic pipeline is to provide historians with tools to analyze 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. Infant mortality is currently defined as a death of a child before reaching the age of one, if subject to age-specific mortality rates of that period. Deaths in the first 24 hours and in the following 27 days 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. Fig. 6 presents the SPARQL query for the deaths in 24 hours after birth. Results of queries are returned in aggregated form without disclosing any identifiable personal data. The death timeframes correspond with specific diseases and whether or not the infant was weaned too early, which can be indicative of lower socio-economic circumstances.

```
SELECT ?s ?DateOfDeath ?AgeInWords {
  ?s a heo:Person.
  ?s heo:AgeAtLastBirthdayInMins ?ageInMins.
  FILTER(?ageInMins <= 1440)
  ?s heo:AgeAtLastBirthday ?AgeInWords.
  ?s heo:dateOfDeath ?DateOfDeath.
}
```

Fig. 6 Example of a use case query for retrieving people who died within 24 hours after birth. Prefixes were omitted and there are 1440 minutes in a day.

7 Longterm Digital Preservation of Data

So far, we have described how the Linked Data platform aids historians in exploring the rich information contained in these vital records and how the interpretation of that information is kept strictly separate from the actual values contained in those records. Interestingly, the efforts of the digital archivists in transcribing these register pages led to the creation of RDF that can be a valuable asset for future research. In this section, we will describe how we will use these transcriptions to create metadata records to store the digitized register pages along with their RDF representations in a suitable longterm digital preservation platform. Some details of this section have been reported in [23].

For the IRL project, we adopted the Digital Repository of Ireland[18] – from now on simply called the Repository – which is a national trusted digital repository for Ireland's social and cultural data. The Repository links together and preserves both historical and contemporary data held by Irish institutions, providing a central internet access point and interactive multimedia tools to the public, students and scholars. Institutions can ingest objects and their metadata one by one via a web-based user interface or in bulk with command line utilities developed for this purpose.

Ingestion of the digitized register pages, metadata records and RDF representations is done as follows:

1. For each register page, we create an RDF file;
2. That RDF file is used to generate a Qualified Dublin Core (QDC) metadata record;
3. All files are ingested into the Repository using the command line utilities.

This process is depicted graphically in Fig. 7 and will be elaborated on in the following subsections.

[18] `http://repository.dri.ie/`

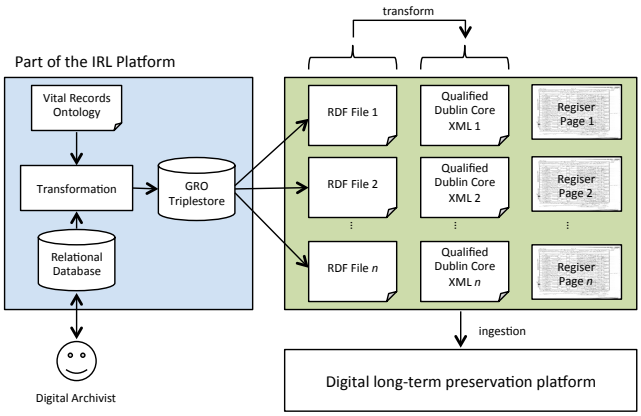


Fig. 7 Generating RDF and Qualified Dublin Core metadata records from the GRO triplestore for longterm digital preservation.

7.1 Creating RDF Files

The transcribed register pages are made available via a SPARQL endpoint. In order to create an RDF document for each register page, we create an RDF model based on a SPARQL DESCRIBE query for each register page's stamp number. An example of such a query is given below (prefixes omitted).

```
DESCRIBE * {
  ?page rec:stampNumber "4646439";
        rec:withRecord ?record.
}
```

This query returns descriptions for all variables in the query; in this case a specific register page and its records. We can write the result to an RDF file, but the file does not state which resource is the "topic" or "subject". To solve this problem, we choose to insert an additional triple that explicitly states that the subject of that file is the register page by using the `foaf:primaryTopic` predicate with the register page's URI. The RDF is serialized as an RDF/XML file using the Stamp ID as the file name.

7.2 Creating Qualified Dublin Core Metadata Records

The guidelines formulated in [13] were aimed at anyone using the Dublin Core metadata standard to prepare content for deposit with the Repository and provides a list of mandatory, recommended and optional fields and, where applicable, suggested controlled vocabularies. In order to create QDC for each register page, we thus have to create and execute a mapping from RDF documents using properties of the Vital Records Ontology to elements in QDC. We adopted XSPARQL [8, 19] to create that mapping. All mandatory fields were mapped and we also covered quite a few of the recommended fields and some optional fields. Note that the RDF does not contain all the information that can or has to be mapped, but constant values can be used. An example of a constant value is attributing copyright, which can be as simple as "Copyright General Register Office Ireland". Most of the register page's information is used to create metadata and each record in the register page is used for a part of the summary in the description field. The result of such a transformation is shown in Listing 1 and the mapping of properties and constant values to QDC fields are shown in Table 1.

7.3 Ingestion into the Repository

The Repository includes a web-based user interface to ingest single objects as well as the facility to ingest metadata and their objects in bulk. For the latter, two directories have to be prepared: metadata and data. The first contains the QDC files – one for each object – and the latter all the digital files associated with the described objects. A file naming convention, described in [13], ensures that the QDC files and digital files are correctly related. The result of bulk ingesting the files into the Repository is shown in Fig. 8, where one can see the metadata and a surrogate of the asset. The Repository provides means to explore both the TIFF as well as the RDF/XML file.

7.4 On the Generated Metadata Records

We have described the three steps for generating QDC metadata records from the RDF files that contain transcriptions of digitized register pages. Indeed, the mapping from the RDF files – which use vocabularies developed for this project – to QDC implies a loss of semantics. However, those metadata records are used to capture the metadata necessary to archive, explore and discover the digitized assets in the Repository. The digitized register page and its transcription using RDF are both stored with that metadata record. Both files can be downloaded from the Repository with the asset browser shown in Fig. 8 (1). The current version of the Repository does not yet support searches within the ingested RDF files, searches are restricted to the content provided in metadata records.

Table 1 Mapping of properties belonging to the Vital Records Ontology used by the RDF resources to Qualified Dublin Core metadata fields for death register pages. Mappings with "as" denote usage of values as is. Mappings with "part of" denote the usage of values as being a part of a field.

Resource	Property/Value	mapped to	QDC Element
Register Page	District	as	Spatial coverage
	Union	as	Spatial coverage
	County	as	Spatial coverage
	Superintendent registrar's district		
	Date certified as true copy by superintendent registrar	as	Date issued
	Date certified by registrar	as	Date created
	Forename/surname registrar on page		
	Forename/surname superintendent registrar	as	creator
	Page number		
	Volume		
	Quarter		
	Stamp number	as	Identifier
		part of	Title
	Year registered	as	Temporal coverage
Record	Date of registration		
	Title/forename/surname registrar		
	Amendments		
	Number in register		
Certificate	Forename/surname (of subject)	part of	Description
	Address (of subject)		
	Sex (of subject)	part of	Description
	Forename/surname informant		
	Qualification of informant		
	Relationship of informant		
	Residence of informant		
Death Record	Forename/surname of registrar		
	Date of death	part of	Description
	Cause of death and duration of illness		
	Condition		
	Age last birthday		
	Place of residence		
	Rank, profession or occupation		
Constant Values	"Copyright General Register Office Ireland"	as	Rights
	"Text"	as	Type
	"en"	as	Language
	"General Register Office"	as	Publisher

8 Considerations whilst Developing the Platform

In this section we elaborate on certain design decisions and aspects of the developed platform and its components.

On Digitized Objects and the Transcriptions. We explained the reason why the Linked Data platform was placed behind a firewall in Section 3. Although not part of this project, one could investigate which subsets of the knowledge bases, and in particular the one containing historical events, do not violate the agreement and could be of benefit to the scientific community. The GRO also digitized the indexes for finding individual records. Indexes are currently not transcribed as they provide no additional information for our data analysis and individual records can be queried with SPARQL.

On Ontology Engineering. The digital archivists keep track of any anomalies or peculiarities in the register pages and individual records in a notes field in the database. Examples of anomalies include strikethroughs in fields or the occurrence of crosses where signatures are necessary. The first *could* indicate a correction or removal of information and the latter could indicate an illiterate person. We carefully chose to use the verb "could" as these are historical vital records and we should not give an interpretation to these anomalies when we are not sure. Depending on the nature of these anomalies and their frequency, we could consider using these for the creation of a controlled vocabulary; allowing one to look up these anomalies and decide how to interpret them. This vocabulary, captured as an ontology, would then reside next to the VRO.

On Qualified Dublin Core. Qualified Dublin Core extends the 15 elements of Simple Dublin Core with 3

Listing 1 The result of transforming RDF into Qualified Dublin Core with XSPARQL. Again with values obfuscated as per our data sharing agreement.

```
<?xml version="1.0" encoding="UTF-8"?>
<qualifieddc xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/
    terms/" xmlns:marcrel="http://www.loc.gov/marc.relators/" xmlns:xsi="http://www.w3.org
    /2001/XMLSchema-instance" xsi:schemaLocation="http://www.loc.gov/marc.relators/ http://
    imlsdcc2.grainger.illinois.edu/registry/marcrel.xsd" xsi:noNamespaceSchemaLocation="
    http://dublincore.org/schemas/xmls/qdc/2008/02/11/qualifieddc.xsd">
   <dc:description>
 Death Register Page 4740277.
 Registered in South City Number 1, South Dublin,
 Dublin in 1890 containing the deaths of
 *******, ***** (M) on 1890-07-18
 ******, ****** (M) on 1890-07-05
 *****, ***** (F) on 1890-07-17
 *****, ********* (F) on 1890-07-15
 **********, ******* (M) on 1890-07-18
 *******, **** (F) on 1890-07-16
 ******, **** (F) on 1890-07-16
 *****, ******** (F) on 1890-07-15
 *****, ***** (M) on 1890-07-18
 ******, **** (F) on 1890-07-16
   </dc:description>
   <dc:rights>Copyright General Register Office Ireland</dc:rights>
   <dc:type>Text</dc:type>
   <dcterms:spatial>South City Number 1, South Dublin, Dublin</dcterms:spatial>
   <dcterms:temporal>1890</dcterms:temporal>
   <dc:language>en</dc:language>
   <dc:identifier>4740277</dc:identifier>
   <dc:publisher>General Register Office</dc:publisher>
</qualifieddc>
```

additional relations and a set of qualifiers that specialize these elements with more specific semantics. Those efforts have then converged into what is known as DCMI Metadata Terms in 2012. The reason we have generated Qualified Dublin Core is that the Digital Repository of Ireland has been developed to ingest Qualified Dublin Core (for which they provided guidelines) and validate those files as such. Other metadata schemas that the repository supports and for which guidelines have been published are Dublin Core [12] and the Metadata Object Description Schema MODS [14]. The former was not chosen, as it was less expressive than Qualified Dublin Core. The latter as the guidelines were only published in February 2016.

9 Related Work

A survey on the application of Semantic Web and Linked Data technologies for historical research was presented in [27] and according to [32], the adoption of Semantic Web technologies in cultural heritage and digital library systems focus on the production of cultural heritage RDF datasets, align these datasets and their vocabularies with external datasets found in the Linked Data cloud, and the exploration and search across data stores.

Though similar practices for ingesting, enriching and preserving metadata exist, such as the Archipel project [15] harvesting data from GLAMS and broadcasters in Flanders (Belgium), we found little related work the transcription, ingestion and preservation of historical vital records:

- [34] proposed a method for extracting information from vital records transcribed as HTML using ontologies. Longterm digital preservation was not an aspect of that study.
- [29] presented an approach to increase the efficiency of identifying potential links across vital records using a person's attributes such as names. Their work is situated in the field of record linking databases.

As for related work in the cultural domain, the authors of [31] reported on a Linked Open Data architecture for the Getuilo Vargas Historical archives. Like our approach, they have adopted semantic technologies, and RDF in particular, as it provides a scalable data model allowing one to more easily build applications for different stakeholders. They have furthermore adopted the provenance ontology PROV-O [26] to rep-

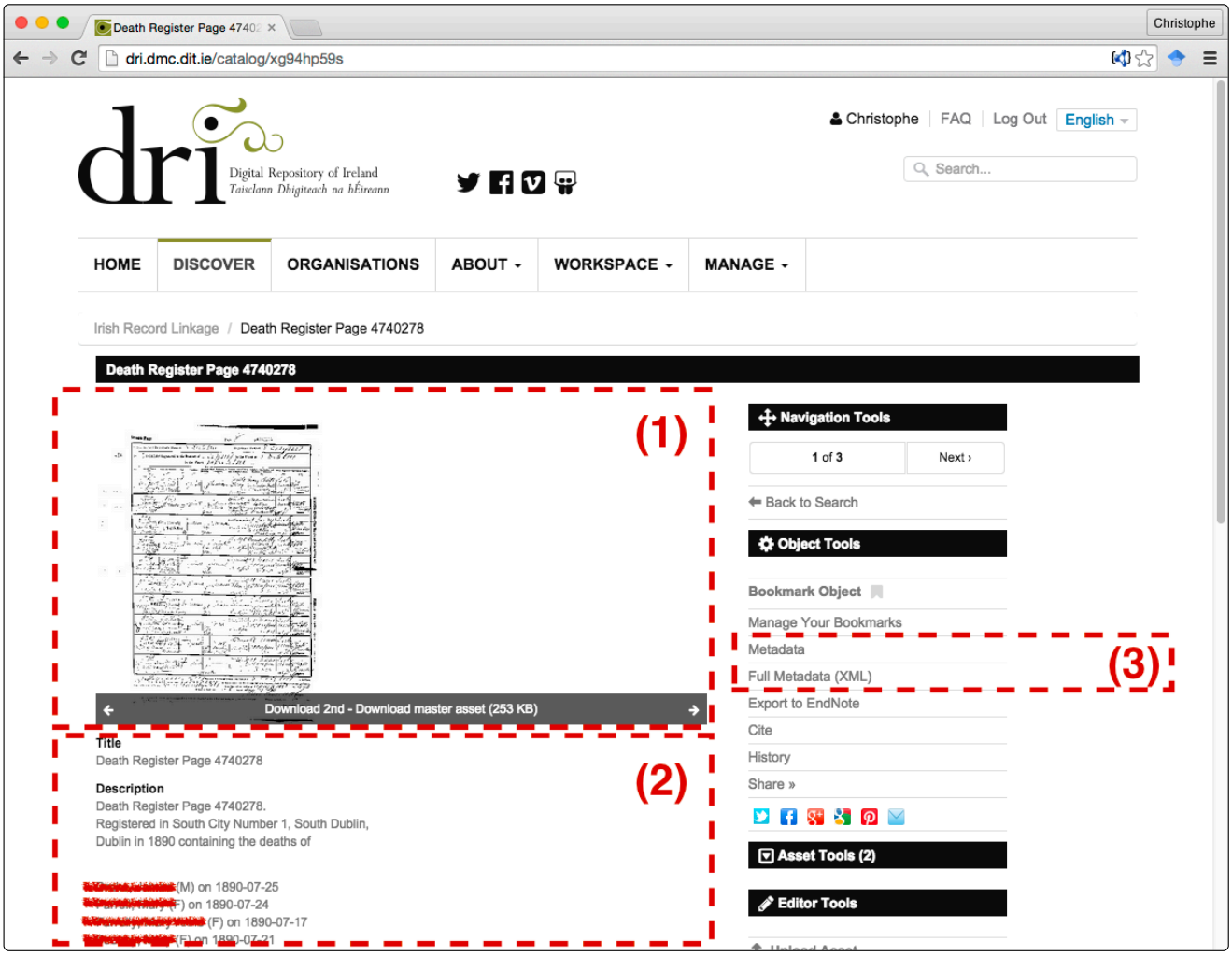


Fig. 8 A register page in the Repository. In (1) we have the assets one can download and for which surrogates are generated. Surrogates are for instance used as thumbnails while browsing collections. In (2), the data provided in the metadata records is shown to the user. The record can also be downloaded as QDC in (3).

resent and exchange provenance information. Interestingly, they also have a manual transcription process for the audio recordings of interviews. Unlike our approach, they have not explicitly modeled the separation of fact and interpretation as part of their methodology.

The provenance of datasets is also a recurring theme of semantics in the cultural heritage domain. [16] reported on PREMIS-OWL, an ontology for capturing the information necessary for longterm preservation of digital assets and their metadata (provenance information, technical information of the digital assets and rights). Their efforts are related and were developed alongside PROV-O, a W3C standard adopted in both [30] and [31]. [30] emphasized the importance of keeping provenance of the generated datasets in data processing "pipelines".

The work presented in [32] mentions interviews with historians for iliciting knowledge which they would use to enrich and complement the information extracted from historical texts. Their work provides an other example of historians and knowledge engineering collaborating to create tools for historians to analyze historical texts. Others looked at linking named entities with external datasets to enable digital humanities scholars. One example is reported in [22], where the authors also examined the results and provide an explanation when their methods (could) fail to create such links.

Although the inclusion of the transcribed and generated datasets in e-infrastructures is outside the scope of this paper, it is worth noting such efforts. An example in the domain of archaeology can be found in the EU funded ARIADNE project that focuses on the integration of archaeological digital resources all over the Europe [2].

10 Conclusions and Future Work

We reported on the creation of the semantic architecture, the ontologies and knowledge bases of the IRL Linked Data platform. Taking into account the requirements of both the digital archivists (archival authenticity, preservation, cataloguing and data protection) and the historians (answering their research questions), the Linked Data platform is comprised of two distinct knowledge bases, each supported by a different ontology, to separate those two concerns: the Vital Records Ontology for the exact transcription of the historical vital records and register pages, and the Historical Events Ontology for an interpretation of the register pages. The creation of the first was fairly straightforward and primarily the result of a collaboration between the knowledge engineers and digital archivists. The latter also involved the historians who were asked to formulate competency questions to identify concepts and relations. Reasoning provides one motivation for adopting semantic technologies. The second is the creation of links with other datasets providing additional context to interpret the data. As the transcription of register pages is a laborious process, the latter can only be meaningfully evaluated when we have an adequate number of transcriptions.

The limitation of the study reported in this paper is the validation of the platform that has been developed by users, which is due to the restricted nature of the data sharing agreement. The terms and conditions of our data sharing agreement did not permit us to make public any data that would identify any individual. Information had to be furthermore obfuscated for all not directly involved in this project and only deployed within the network of the Royal Irish Academy. Meaningful user trials and experiments involving users were therefore not feasible. Other components, such as the Digital Repository we have adopted, have been trialed outside this project and as for the mappings from RDF to QDC, the librarians and archivists involved in this project – 2 people – have looked at and provided feedback on the QDC that had been generated.

The lessons learned in this study arise from the value of the separation of concerns. Though the digital archivists could have elicited facts from the register pages immediately and solely fit for answering the competency questions in this project, the resulting dataset would have had limited value for reuse and future research questions. We argue that the return in value justified the extra overhead in terms of transcription and platform complexity. Our approach is thus different from, for instance, the Dacura platform [20], which adopts crowdsourcing techniques to elicit facts from datasets such as newspaper articles according to a schema for a particular purpose.

Another valuable lesson is the collaboration between digital archivists and knowledge (Linked Data) engineers in developing this platform. Both roles come with different skill sets and perspectives in capturing information and knowledge. Informed decisions were made whilst developing the platform – e.g., digital archivists providing information to the knowledge engineers on how anomalies can be captured and knowledge engineers helping digital archivists "normalize" the information – that helped us gather information beyond what is available in the forms provided by the register pages.

References

1. First Annual Report of the Registrar-General of Marriages, Births, and Deaths in Ireland. (1869). URL `http://www.cso.ie/en/media/csoie/releasespublications/documents/birthsdm/archivedreports/P-VS,1864.pdf`
2. Aloia, N., Papatheodorou, C., Gavrilis, D., Debole, F., Meghini, C.: Describing research data: A case study for archaeology. In: R. Meersman, H. Panetto, T.S. Dillon, M. Missikoff, L. Liu, O. Pastor, A. Cuzzocrea, T.K. Sellis (eds.) On the Move to Meaningful Internet Systems: OTM 2014 Conferences - Confederated International Conferences: CoopIS, and ODBASE 2014, Amantea, Italy, October 27-31, 2014, Proceedings, *Lecture Notes in Computer Science*, vol. 8841, pp. 768–775. Springer (2014). DOI 10.1007/978-3-662-45563-0_48
3. Arenas, M., Bertails, A., Prudhommeaux, E., Sequeda, J.: A Direct Mapping of Relational Data to RDF. W3C Recommendation, W3C (2012). URL `https://www.w3.org/TR/rdb-direct-mapping/`
4. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: A Nucleus for a Web of Open Data. In: K. Aberer, K. Choi, N.F. Noy, D. Allemang, K. Lee, L.J.B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, P. Cudré-Mauroux (eds.) The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007., *Lecture Notes in Computer Science*, vol. 4825, pp. 722–735. Springer (2007)
5. Berners-Lee, T.: Relational databases on the semantic web. http://www.w3.org/DesignIssues/RDB-RDF.html (last retrieved December 2012) (1998)
6. Berners-Lee, T.: Linked Data - Design Issues. Last accessed: June 7th, 2015. (2006). URL `http://www.w3.org/DesignIssues/LinkedData.html`
7. Beyan, O., Breathnach, C., Collins, S., Debruyne, C., Decker, S., Grant, D., Grant, R., Gurrin, B.: Towards Linked Vital Registration Data for Reconstituting Families and Creating Longitudinal Health Histories. In: KR4HC Workshop (in conjunction with KR 2014), pp. 181–187 (2014)
8. Bischof, S., Decker, S., Krennwallner, T., Lopes, N., Polleres, A.: Mapping between RDF and XML with XSPARQL. J. Data Semantics **1**(3), 147–185 (2012)
9. Bizer, C.: D2R MAP - A database to RDF mapping language. In: I. King, T. Máray (eds.) Proceedings of

the Twelfth International World Wide Web Conference - Posters, WWW 2003, Budapest, Hungary, May 20-24, 2003 (2003)
10. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. Int. J. Semantic Web Inf. Syst. **5**(3), 1–22 (2009)
11. Boonstra, O., Breure, L., Doorn, P.: Past, present and future of historical information science. Historical Social Research/Historische Sozialforschung pp. 4–132 (2004)
12. Bustillo, M., Collins, S., Gallagher, D., Grant, R., Harrower, N., Kenny, S., Ní Cholla, R., O'Carroll, A., Redmond, S., Webb, S.: Dublin Core and the Digital Repository of Ireland (Grant, R. ed.). Tech. rep., Maynooth: Maynooth University; Dublin: Trinity College Dublin; Dublin: Royal Irish Academy; Galway: National University of Ireland, Galway (2014)
13. Bustillo, M., Collins, S., Gallagher, D., Grant, R., Harrower, N., Kenny, S., Ní Cholla, R., O'Carroll, A., Redmond, S., Webb, S.: Qualified Dublin Core and the Digital Repository of Ireland (Grant, R. ed.). Tech. rep., Maynooth: Maynooth University; Dublin: Trinity College Dublin; Dublin: Royal Irish Academy; Galway: National University of Ireland, Galway (2015)
14. Bustillo, M., Grant, R., Kenny, S., Martínez-García, A., McGoohan, C., Ní Cholla, R., O'Carroll, A., O'Neill, J., Redmond, S., Webb, S.: MODS and the Digital Repository of Ireland (Grant, R. ed.). Tech. rep., Maynooth: Maynooth University; Dublin: Trinity College Dublin; Dublin: Royal Irish Academy; Galway: National University of Ireland, Galway (2016)
15. Coppens, S., Mannens, E., Van Deursen, D., Hochstenbach, P., Janssens, B., Van de Walle, R.: Publishing provenance information on the web using the memento datetime content negotiation. In: C. Bizer, T. Heath, T. Berners-Lee, M. Hausenblas (eds.) WWW2011 Workshop on Linked Data on the Web, Hyderabad, India, March 29, 2011, *CEUR Workshop Proceedings*, vol. 813. CEUR-WS.org (2011)
16. Coppens, S., Verborgh, R., Peyrard, S., Ford, K., Creighton, T., Guenther, R., Mannens, E., Van de Walle, R.: PREMIS OWL - A semantic long-term preservation model. Int. J. on Digital Libraries **15**(2-4), 87–101 (2015). DOI 10.1007/s00799-014-0136-9
17. Das, S., Sundara, S., Cyganiak, R.: R2RML: RDB to RDF Mapping Language. W3C Recommendation, W3C (2012). URL `http://www.w3.org/TR/r2rml/`
18. Debruyne, C., Beyan, O.D., Grant, R., Collins, S., Decker, S.: On a linked data platform for irish historical vital records. In: S. Kapidakis, C. Mazurek, M. Werla (eds.) Research and Advanced Technology for Digital Libraries - 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015. Proceedings, *Lecture Notes in Computer Science*, vol. 9316, pp. 99–110. Springer (2015). DOI 10.1007/978-3-319-24592-8_8
19. Dell'Aglio, D., Polleres, A., Lopes, N., Bischof, S.: Querying the web of data with XSPARQL 1.1. In: R. Verborgh, E. Mannens (eds.) Proceedings of the ISWC Developers Workshop 2014, co-located with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 19, 2014., *CEUR Workshop Proceedings*, vol. 1268, pp. 113–118. CEUR-WS.org (2014)
20. Feeney, K.C., O'Sullivan, D., Tai, W., Brennan, R.: Improving curated web-data quality with structured harvesting and assessment. Int. J. Semantic Web Inf. Syst. **10**(2), 35–62 (2014). DOI 10.4018/ijswis.2014040103
21. Fox, M.S., Gruninger, M.: Enterprise modeling. AI magazine **19**(3), 109–121 (1998)
22. Frontini, F., Brando, C., Ganascia, J.: Semantic web based named entity linking for digital humanities and heritage texts. In: Zucker et al. [35], pp. 77–88. URL `http://ceur-ws.org/Vol-1364/paper9.pdf`
23. Grant, D., Debruyne, C., Grant, R., Collins, S.: Creating and consuming metadata from transcribed historical vital records for ingestion in a long-term digital preservation platform. In: I. Ciuciu, H. Panetto, C. Debruyne, A. Aubry, P. Bollen, R. Valencia-García, A. Mishra, A. Fensel, F. Ferri (eds.) On the Move to Meaningful Internet Systems: OTM 2015 Workshops, *Lecture Notes in Computer Science*, vol. 9416, pp. 445–450. Springer (2015). DOI 10.1007/978-3-319-26138-6_47
24. Grüninger, M., Fox, M.S.: The role of competency questions in enterprise engineering. In: BenchmarkingTheory and Practice, pp. 22–31. Springer (1995)
25. Lopes, N., Grant, R., Ó Raghallaigh, B., Ó Carragáin, E., Collins, S., Decker, S.: Linked Logainm: Enhancing Library Metadata Using Linked Data of Irish Place Names. In: L. Bolikowski, V. Casarosa, P. Goodale, N. Houssos, P. Manghi, J. Schirrwagen (eds.) Theory and Practice of Digital Libraries - TPDL 2013 Selected Workshops - LCPD 2013, SUEDL 2013, DataCur 2013, Held in Valletta, Malta, September 22-26, 2013. Revised Selected Papers, *Communications in Computer and Information Science*, vol. 416, pp. 65–76. Springer (2014)
26. McGuinness, D., Lebo, T., Sahoo, S.: PROV-O: The PROV ontology. W3C Recommendation, W3C (2013). Http://www.w3.org/TR/2013/REC-prov-o-20130430/
27. Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., van Harmelen, F.: Semantic technologies for historical research: A survey. Semantic Web **6**(6), 539–564 (2015). DOI 10.3233/SW-140158. URL `http://dx.doi.org/10.3233/SW-140158`
28. Motik, B., Horrocks, I.: OWL Datatypes: Design and Implementation. In: A.P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T.W. Finin, K. Thirunarayan (eds.) The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings, *Lecture Notes in Computer Science*, vol. 5318, pp. 307–322. Springer (2008)
29. Newcombe, H.B., Kennedy, J.M.: Record linkage: Making maximum use of the discriminating power of identifying information. Commun. ACM **5**(11), 563–566 (1962). DOI 10.1145/368996.369026
30. Orgel, T., Höffernig, M., Bailer, W., Russegger, S.: A metadata model and mapping approach for facilitating access to heterogeneous cultural heritage assets. Int. J. on Digital Libraries **15**(2-4), 189–207 (2015). DOI 10.1007/s00799-015-0138-2
31. Rademaker, A., Oliveira, D., de Paiva, V., Higuchi, S., Sá, A., Alvim, M.: A linked open data architecture for the historical archives of the getulio vargas foundation. Int. J. on Digital Libraries **15**(2-4), 153–167 (2015). DOI 10.1007/s00799-015-0147-1
32. Tounsi, M., Faron-Zucker, C., Zucker, A., Villata, S., Cabrio, E.: Studying the history of pre-modern zoology with linked data and vocabularies. In: Zucker et al. [35], pp. 7–14. URL `http://ceur-ws.org/Vol-1364/paper1.pdf`
33. Unbehauen, J., Hellmann, S., Auer, S., Stadler, C.: Knowledge extraction from structured sources. In:

S. Ceri, M. Brambilla (eds.) Search Computing - Broadening Web Search, *Lecture Notes in Computer Science*, vol. 7538, pp. 34–52. Springer (2012). DOI 10.1007/978-3-642-34213-4_3. URL http://dx.doi.org/10.1007/978-3-642-34213-4_3

34. Woodbury, C.: Automatic extraction from and reasoning about genealogical records: A prototype. Master's thesis, Brigham Young University, Provo, Utah, United States (2010)
35. Zucker, A., Draelants, I., Faron-Zucker, C., Monnin, A. (eds.): Proceedings of the First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference, Portorož, Slovenia, June 1st, 2015, *CEUR Workshop Proceedings*, vol. 1364. CEUR-WS.org (2015). URL http://ceur-ws.org/Vol-1364