Dolores Grant, Christophe Debruyne, Rebecca Grant, Sandra Collins: Creating and Consuming Metadata from Transcribed Historical Vital Records for Ingestion in a Long-Term Digital Preservation Platform - (Short Paper). OTM Workshops 2015: 445-450
The final version is available at http://link.springer.com/
http://link.springer.com/chapter/10.1007%2F978-3-319-26138-6_47

Creating and Consuming Metadata from Transcribed Historical Vital Records for Ingestion in a Long-term Digital Preservation Platform (short paper)

Dolores Grant[1], Christophe Debruyne[2,3], Rebecca Grant[1], and Sandra Collins[1]

[1] Digital Repository of Ireland, Royal Irish Academy, Dublin 2, Ireland
{d.grant,r.grant,s.collins}@ria.ie
[2] ADAPT Centre for Digital Content Platform Research, Knowledge & Data Engineering Group, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland
debruync@scss.tcd.ie
[3] Web & Information Systems Engineering Laboratory, Department of Computer Science, Vrije Universiteit Brussel, Brussels, Belgium
chrdebru@vub.ac.be

Abstract. In the Irish Record Linkage 1864-1913 (IRL) project, digital archivists transcribe digitized register pages containing vital records into a database, which is then used to generate RDF triples. Historians then use those triples to answer some specific research questions on the IRL platform. Though the triples themselves are a highly valuable asset that can be adopted by many, the digitized records and their RDF representations need to be adequately stored and preserved according to best standards and guidelines to ensure those do not get lost over time. This was a problem currently not investigated within this project. This paper reports on the creation of Qualified Dublin Core from those triples for ingestion with the digitized register pages in an adequate long-term digital preservation platform and repository. Rather than creating RDF only for the purpose of this project, we demonstrate how we can distill artifacts from the RDF that is fit for discovery, access, and even reuse via that repository and how we elicit and conserve the knowledge and memories about Ireland, its history and culture contained in those register pages.

Keywords. Linked Data, Metadata, Mapping, Vital Records

1 Introduction

The IRL[1] project aims to create a knowledge base containing historical birth-, marriage- and death records translated into RDF and to create a Linked Data platform to analyze those events. In [1], we reported on the semantic architecture in which we separate two concerns: 1) the exact transcription of the register pages from TIFF files (provided by the General Register Office (GRO), Ireland's central civil repository for

[1] Irish Record Linkage, 1864-1913: https://irishrecordlinkage.wordpress.com/

records relating to births, marriages and deaths in Ireland) by the digital archivists transformed into RDF using the Vital Records Ontology[2] and 2) the interpretation thereof by the historians. This requirement resulted in a platform with two distinct knowledge bases where the interpretation refers back to the transcribed register pages, but the knowledge base containing those transcribed register pages cannot be "contaminated" with any other knowledge.

In this paper, we report on the creation of metadata records from the generated RDF to facilitate the exploration of register pages and vital records in an adequate long-term digital preservation platform. We will describe our method to distil Qualified Dublin Core metadata records for each register page and how these can be ingested together with the TIFF and RDF representation. We must note that the terms and conditions of our data sharing agreement with the GRO do not permit us to make public any data that would identify any individual. We will thus obfuscate information concerning individuals where necessary.

2 Related Projects and Initiatives

Though similar practices for ingesting, enriching and preserving metadata exist, such as the Archipel project [5] harvesting data from GLAMS and broadcasters in Flanders (Belgium), we found little related work the transcription, ingestion and preservation of historical vital records. A method for extracting information from vital records transcribed as HTML using ontologies was proposed in [7]. Long-term digital preservation was not an aspect of that study. [6] presented an approach to increase the efficiency of identifying potential links across vital records based on a person's attributes such as names. Their work is situated in the field of record linking databases.

3 Method

The creation of RDF from the transcribed register pages using the Vital Records Ontology was reported in [1]. The creation of the metadata for ingestion into the Digital Repository of Ireland, from now on called the Repository, will be described in this section. The Repository allows one to ingest metadata and related assets in bulk. We adopted the guidelines in [2] for the creation of our Qualified Dublin Core (QDC) metadata records for each register page. We then prepared an RDF file and retrieved the digital surrogate of those register pages (in TIFF format) in such a way that they will be associated with their corresponding QDC file during ingestion.

3.1 DRI Bulk Ingestion

Though technical and not exactly relevant for this paper, we feel it is important to elaborate on the bulk ingestion facilities of 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 ensures that the QDC files and digital files are correctly related.

[2] `http://www.purl.org/net/irish-record-linkage/records`

The result of bulk ingesting the files into the Repository is shown in **Fig. 1**, 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. Before one can ingest the files, both the metadata record and RDF/XML files need to be generated. This process will be described in the following sections.

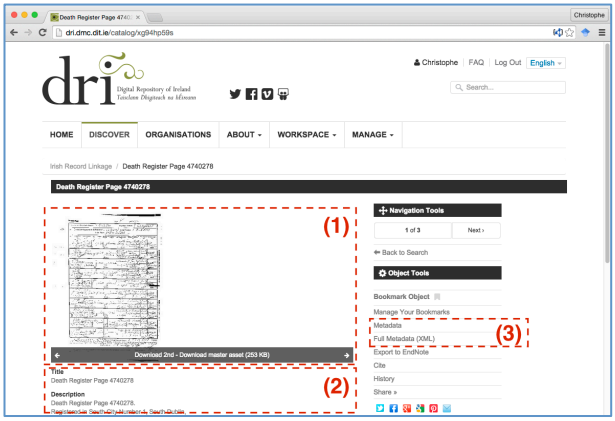

Fig. 1. A register page in the Repository. In (1) we have the assets one can download and for which surrogates are generated. Surrogates are for instance used as thumbnails while browsing collections. In (2), the data provided in the metadata records is shown to the user. The record can also be downloaded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

3.2 Creation of RDF for Register Pages

The transcribed register pages are transformed into RDF, loaded into a triplestore and 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.

```
PREFIX rec: <http://purl.org/net/irish-record-linkage/records#>
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" of "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 file is written to the data directory's folder with the Stamp ID as the file name. This folder also contains the digitized register pages with the same name.

3.3 The Creation of Qualified Dublin Core Metadata Records

The guidelines formulated in [5] were aimed at anyone using the Dublin Core metadata standard to prepare content for deposition 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 the RDF to elements in QDC. We adopted XSPARQL [5] 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 **Fig. 2**.

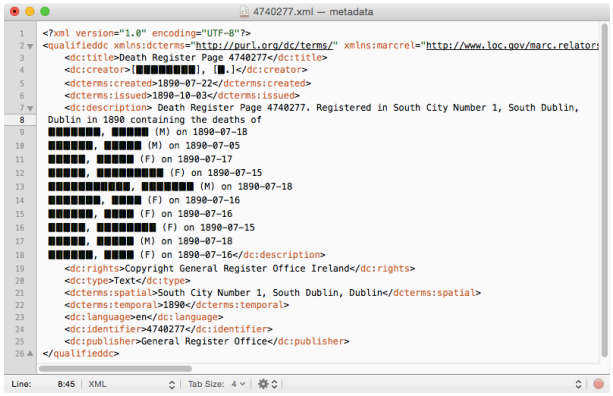

Fig. 2. The result of transforming RDF into Qualified Dublin Core with XSPARQL.

4 Discussion

The work presented in this paper focused on the creation of suitable a mapping from RDF to Qualified Dublin Core for ingestion into a long-term digital preservation platform. One limitation of this study is investigating to what extent the metadata records are adequate for one to find and discover those records not only from the perspective of cataloguers and archivist, but also from the perspective of end users. This limitation is largely due to the sensitive nature of the data, which we hope to address in the future. Due to the sensitive nature of the IRL data, we were also unable to adopt any crowdsourcing mechanisms to leverage the transcription process. Currently, two digital archivists are transcribing the records and quality checking each other's output.

5 Conclusions and Future Work

In the Irish Record Linkage 1864-1913 (IRL) project, digital archivists transcribe historical vital records in register pages, which are transformed into RDF with the Vital Records Ontology. This means that those records are available both as TIFF and RDF. What has not yet been investigated in the project, however, is how these files can be ingested in adequate long-term digital preservation platforms to ensure that this rich information does not get lost. In this paper, we reported on the process of creating RDF files for each register page followed by the creation of a Qualified Dublin Core (QDC) metadata record according to best practices, standards and guidelines.

For each register page, we ingested the scan, an RDF file and a DQC file into the Digital Repository of Ireland. We thus demonstrated how the RDF generated in the IRL project was reused to create other structured data that allows one to discover and reuse the information captured in those register pages.

A limitation of this study is the lack of investigating to what extent the mapping of RDF to QDC generates adequate metadata records from a cataloguing perspective and evaluating the to what extent the information in the QDC we generated is rich enough for users to explore. Finally, we are currently investigating the adoption of Encoded Archival Description (EAD) to catalog the register pages, the records as parts of those register pages and the database currently being populated by the digital archivists. The results of this exercise, as well as a comparison with the metadata in QDC is will be reported elsewhere.

Acknowledgements. We thank the Registrar General of Ireland for permitting us to use the vital records for the purposes of this research project. This publication has emanated from research conducted within the Irish Record Linkage, 1864-1913 project supported by the RPG2013-3; Irish Research Council Interdisciplinary Research Project Grant, and within the Science Foundation Ireland Funded Insight Research Centre (SFI/12/RC/2289). The Digital Repository of Ireland (formerly NAVR) acknowledges funding from the Irish HEA PRTLI programme. Christophe Debruyne is supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of the ADAPT Centre for Digital Content Platform Research at Trinity College Dublin.

6 References

1. 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)
2. 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)
3. Dell'Aglio, D., Polleres, A., Lopes, N., Bischof, S.: Querying the Web of Data with XSPARQL 1.1. In: Verborgh, R., Mannens, E. (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 Work- shop Proceedings, vol. 1268, pp. 113–118. CEUR-WS.org (2014)
4. Harris, S., Seaborne, A.: SPARQL 1.1 query language. W3C Recommendation, W3C (Mar 2013), http://www.w3.org/TR/2013/REC-sparql11-query-20130321/
5. Coppens, S., Mannens, E., Deursen, D.V., Hochstenbach, P., Janssens, B., de Walle, R.V.: Publishing provenance information on the web using the memento date- time content negotiation. In: Bizer, C., Heath, T., Berners-Lee, T., Hausenblas, M. (eds.) WWW2011 Workshop on Linked Data on the Web, Hyderabad, India, March 29, 2011. CEUR Workshop Proceedings, vol. 813. CEUR-WS.org (2011)
6. Newcombe, H.B., Kennedy, J.M.: Record linkage: making maximum use of the discriminating power of identifying information. Communication of ACM 5(11), 563–566 (1962)
7. Woodbury, C.: Automatic extraction from and reasoning about genealogical records: A prototype. Master's thesis, Brigham Young University (2010)