Creating a Knowledge Graph for Ireland's Lost History: Knowledge Engineering and Curation in the Beyond 2022 Project

CHRISTOPHE DEBRUYNE and GARY MUNNELLY, ADAPT Centre, Trinity College Dublin
LYNN KILGALLON, Department of History, Trinity College Dublin
DECLAN O'SULLIVAN, ADAPT Centre, Trinity College Dublin
PETER CROOKS, Department of History, Trinity College Dublin

The Beyond 2022 project aims to create a virtual archive by digitally reconstructing and digitizing historical records lost in a catastrophic fire which consumed items in the Public Record Office of Ireland in 1922. The project is developing a knowledge graph (KG) to facilitate information retrieval and discovery over the reconstructed items. The project decided to adopt Semantic Web technologies to support its distributed KG and reasoning. In this article, we present our approach to KG generation and management. We elaborate on how we help historians contribute to the KG (via a suite of spreadsheets) and its ontology. We furthermore demonstrate how we use named graphs to store different versions of factoids and their provenance information and how these are serviced in two different endpoints. Modeling data in this manner allows us to acknowledge that history is, to some extent, subjective and different perspectives can exist in parallel. The construction of the KG is driven by competency questions elicited from subject matter experts within the consortium. We avail of CIDOC-CRM as our KG's foundation, though we needed to extend this ontology with various qualifiers (types) and relations to support the competency questions. We illustrate how one can explore the KG to gain insights and answer questions. We conclude that CIDOC-CRM provides an adequate, albeit complex, foundation for the KG and that named graphs and Linked Data principles are a suitable mechanism to manage sets of factoids and their provenance.

CCS Concepts: • **Computing methodologies** → **Ontology engineering**; • **Information systems** → **Mediators and data integration**; • **Applied computing** → **Arts and humanities**;

Additional Key Words and Phrases: Knowledge graph creation, knowledge graph management, digital humanities

25

Authors' addresses: C. Debruyne, G. Munnelly, and D. O'Sullivan, ADAPT Centre, Trinity College Dublin, College Green, Dublin 2, Ireland; emails: {debruync, munnellg, declan.osullivan}@tcd.ie; L. Kilgallon and P. Crooks, Department of History, Trinity College Dublin, College Green, Dublin 2, Ireland; emails: {kilgalll, pcrooks}@tcd.ie.
Authors current address: C. Debruyne's, Montefiore Institute, University of Liège, Quartier Polytech 1, Allée de la découverte 10, 4000 Liège, Belgium.

1 INTRODUCTION

On June 30th, 1922, at the beginning of the Irish civil war, the western block of the Four Courts, Dublin was hit by a terrible explosion, ignited during a siege between anti-treaty forces and the Provisional Government of Ireland [9]. The resulting fire destroyed the **Public Record Office of Ireland** (**PROI**) and, with it, seven centuries of Ireland's collective memories. Across the globe, more than 70 repositories hold substitute materials that can replace the lost documents. The Beyond 2022[1] project aims to assemble a complete inventory of loss and survival of the contents with the goal of virtually recreating materials in the PROI of the Four Courts. The project and the resulting virtual treasury will gather all the information it can about these substitute sources from archives and libraries in Ireland and internationally.

Within Beyond 2022, there are four pillars representing strands of scholarly activity involved in assembling the virtual record treasury: Discover, Digitize, Reconstruct, and Reveal. The Reveal strand incorporates research into techniques that facilitate information retrieval, discovery, and enhanced accessibility to the contents of the virtual record treasury. In this article, we will focus on the **knowledge graph** (**KG**) construction for the Beyond 2022 project, which is a core undertaking in this research strand. We will also discuss the ontologies adopted (and developed) to help answer these competency questions. The project's KG gathers information from both subject matter experts (e.g., historians collecting information) as well as automated techniques (e.g., named entity recognition in digitized documents). As subject matter experts may approach the information with a certain bias (e.g., background knowledge or theory) and computer-based agents may use different algorithms and even different parameters, provenance will be vital to ensuring that a user can evaluate themselves the authoritativeness of different parts of the KG. To support this, we designed an interdisciplinary approach to historical data management using KG technologies. The use of named graphs is key to our approach to representing the different interpretations and mutable nature of historical knowledge.

The contributions of this article are, therefore, the *interdisciplinary* approach to KG management, our methodology to KG generation, and the demonstration of our approach. Our contribution is thus particularly relevant for interdisciplinary digital humanities projects in which KG technologies are used to integrate, share, and process information in a flexible and meaningful manner. Interdisciplinary work of this kind needs to draw upon the skills and competencies of both historians and knowledge engineers and thus requires close collaboration between them at every stage of the KG's conceptualization, design and development. Section 2 describes the considerations that informed our approach to knowledge organization (Section 3), ontology engineering (Section 4), KG construction (Section 5), and management (Section 6). In Section 7, we describe a case study: creating RDF from entities mentioned in the Irish Exchequer Index. We describe related work in Section 8 and conclude the article in Section 9.

2 KNOWLEDGE GRAPH DESIGN CONSIDERATIONS

At its core, Beyond 2022 is a reconstruction project involving the assembly of numerous historical resources across a multitude of archives to reconstitute the lost content of destroyed documents. The information about reconstruction has been captured in a relational database, and, for the foreseeable future, this practice will continue. While the relational database's purpose is to store information about a lost item and its substitute, the KG's purpose is to model information contained within the reconstructed sources such as books and parchments (see [30]), not information about the reconstruction itself. Entities in the KG will refer to entries in the database. The KG will thus integrate information from both the sources that were lost or damaged in the fire and additional material to describe and contextualize the entities. An example of the latter is the *Irish Exchequer Payments, 1270-1446* [2] which will be described in our case study in Section 7.
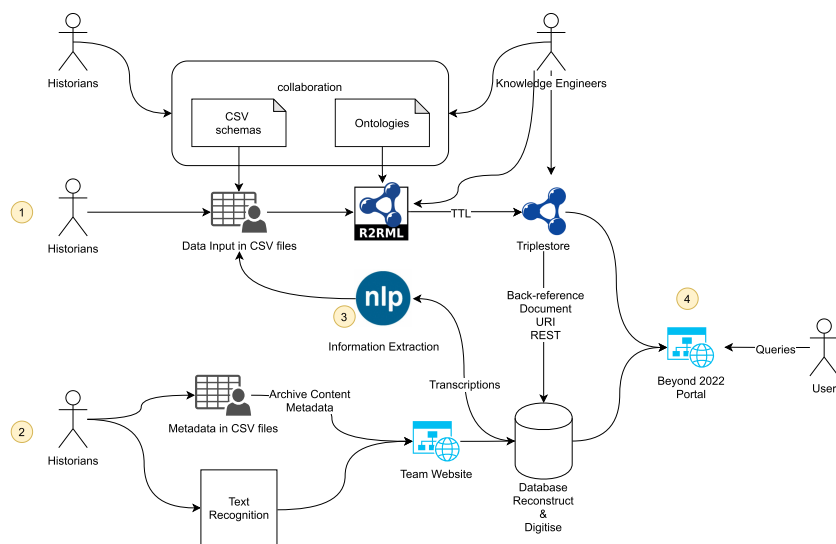
---

[1]https://beyond2022.ie/.

Fig. 1. This diagram captures the role of the NLP pipeline and the KG. In Beyond 2022, the first aim is to create metadata about archives destroyed in the fire and to digitize substitute materials in other repositories (2). The digitized assets are processed using the Handwritten Text Recognition software developed by Transkribus [1] in order to make the content of the assets text-searchable. The project's second aim is to create a KG of entities mentioned within all those digitized assets (1). The relationships, scope, and granularity go beyond what is contained in these assets (e.g., via completion with additional research). Within the project, the NLP pipeline will extract and reconcile entities within documents to populate the spreadsheets (input schemas), which will then be transformed into triples for the KG. Historians and knowledge engineer collaborate on the ontology and CSV schemas. Knowledge engineer are responsible for the R2RML mappings to transform CSV files into RDF and the knowledge organization and knowledge management of the KG.

Information is added to the KG by two means: a manual aggregation of data performed by historians and (more experimentally) by the automatic extraction process performed by a **Natural Language Processing** (**NLP**) pipeline. In this article, we will focus on the former, though it has influenced the design of our work-flow and knowledge organization strategy. The overarching goal is depicted in Figure 1. The project aims to digitize physical archives and make the digital surrogates, together with text-searchable content and metadata, available in a document repository. Simultaneously, historians are conducting research and populating various data schemas, which are then transformed into a KG (2). It is important to note that we consider the digitized as-sets separate from the entities mentioned in those documents. Indeed, the KG may capture factoids[2] beyond the digitized assets, and even in a more granular fashion. These digitized items are, however, considered to populate the KG via an NLP pipeline, of which the output will populate the same data schemas that the historians use (3). How the factoids gathered from the NLP process are kept separate from those curated by the historians is part of our knowledge organization strategy that we will explain later on. The KG does link back to the digitized assets when historians or the NLP pipeline capture those references in the schema. Finally, the user accessing the portal can either search within the documents database or the KG, and can avail of links to switch from one to the other.

---

[2]A note on our use of the term "factoid". It is generally accepted in Beyond 2022 that history is open to interpretation, and the information which populates the KG represents only a single perspective of the past. We use "factoid" rather than "fact" to describe information repre-sented by our triples in order to emphasize that we are not declaring our information to be definitively true. We acknowledge that anything we model is subject to debate, and welcome this as a possibility.

Given this context, the considerations are as follows:

(1) Be it the interpretation of a historian, information contained in a source, or the results of a heuristic, the project should take into account the provenance of so-called factoids. This provenance information is important for a user to assess the quality and trustworthiness of a set of factoids. In Section 3, we elaborate on how the **Resource Description Framework** (**RDF**) [29] and the SPARQL query language for RDF [15] not only provides us a distributed KG, but also facilitates separating those sets of factoids into "containers" with support for tracking its evolution over time. The use of these containers (called named graphs in RDF) and provenance information would allow one to assess the information gathered by those manual and (semi-)automatic approaches.

(2) Ontologies are commonly defined as "a [formal] explicit specification of a [shared] conceptualization" [13]. Ontologies capture a group of stakeholders' shared understanding of a **Universe of Discourse** (**UoD**), either through collaboration or by sharing sufficiently detailed documentation. Within the Semantic Web community, some consider an ontology the set of concepts, relations, instances, and relationships.[3] Others, such as [17], consider the ontology a "schema" that holds in multiple scenarios. In other words, the ontology contains the concepts, relations, and common individuals shared across different universes of discourse. We follow [17] in that the Beyond 2022 ontology will contain concepts, relations, and instances[4] that will be used to represent entities in various sources. The combination of an ontology and a set of instances and relationships is then called a **knowledge base** (**KB**). A KG is a set of interconnected typed entities and their attributes and relationships [12]. Ontologies provide the types and relations. While confusing, it suffices to state that all KGs are KBs, but the converse is not necessarily true. For a KB to be considered a KG, a KB has to use graph technologies such as RDF.

While we know what an ontology is, a challenge within the Beyond 2022 project is to construct its ontology. We adopted an international standard for the cultural heritage domain called the **CIDOC Conceptual Reference Model** (**CIDOC-CRM**)[5] heritage domain but too generic on its own for our purposes. The ontology has to be extended with bespoke concepts. We support two approaches: a top-down approach in which historians structure their thoughts using a controlled natural language (subsequently "implemented" by knowledge engineers), and a bottom-up approach in which historians can introduce new concepts when entering factoids for subsequent consideration to be included in the ontology. We elaborate on these approaches in Section 4.

(3) Historians should be able to add data to the KG in an intuitive manner. We believe that we have the greatest chance of engaging historians to curate information in our KG if we enable them to work with the tools that are already familiar to them. This led us to design our data capture process around the use of prescribed spreadsheets centered around key concepts (People, Office, Organization, and Place), which set out the information that should be captured by a historian seeking to add information to the KG. Knowledge Engineers then created mappings to generate RDF from these spreadsheets, hiding the complexity from the historians. The approach is scalable, as expanding our data capture process to encompass new entity types is simply a matter of designing a new spreadsheet and developing a mapping between its columns and the KG. This is discussed in Section 5.

(4) Historians should be able to scrutinize the contents of the KG, assess the quality of its contents, and modify triples where they are found to be erroneous. As fallible humans, errors will undoubtedly occur in the spreadsheets. It should be possible for historians to check the triples generated by the content they have uploaded to ensure that no mistakes have been made. We discuss this in Section 6.

---

[3]Here, a relationship is an instance of a relation; e.g., the relationship "Christophe knows Lynn" is an instance of the relation "knows".
[4][17] calls instances that are shared across UoDs "ontology individuals".
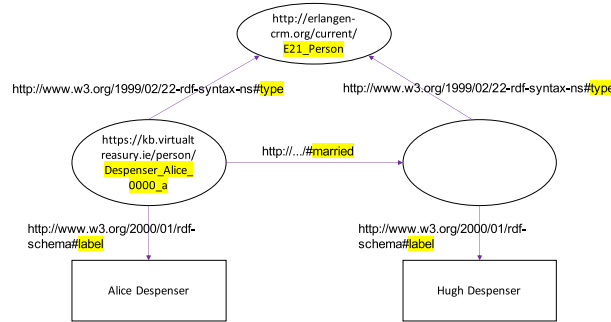[5]http://www.cidoc-crm.org/.

Fig. 2. RDF graph containing five RDF statements. The example represents two resources that represent people. They are "tagged" as such via the statements "type" of "E21_Person". A IRI identifies one resource, and the other is a blank node. Both have a literal attached to via the "label" predicate. Finally, this example also captures how the first person is married to the second. All but the marriage statement comes from the Beyond 2022 KG. Marriages are represented as events, which are a bit more complicated and would have defeated the purpose of illustrating the RDF data model.

3 KNOWLEDGE ORGANIZATION AND PROVENANCE

3.1 Resource Description Framework

The knowledge in the KG is represented using the RDF standard. Figure 2 graphically depicts such a graph. In RDF, "things" are either identified by a IRI[6] or a literal. The difference between the two is that the former identifies and refers to "things" that cannot be printed on a screen (such as a person), and the latter refers to things that can (such as a person's name). Both IRIs and literals are considered *resources* in the RDF model. RDF also provides a means to describe things for which there is no identifier called blank nodes. A blank node is used to declare the existence of something without having a reference for it. We use blank nodes for people that are of no particular interest or for which we do not have sufficient information. It happens that people are mentioned "in passing" (e.g., "the wife of Hugh Despenser"). It is up to the historian to decide whether that person warrants a resource with an identifier. When they deem the information is merely contextual, then a blank node suffices.

Without going into too much technical detail, suffice to say that an RDF graph comprises a set of RDF statements. In the Beyond 2022 project, we store the factoids as such statements. Each RDF statement, called a triple, consists of a subject, a predicate, and an object. A subject is always a IRI or a blank node, the predicate is always a IRI, and the object can be either a IRI, a blank node, or a literal. One can see that the subject of one RDF statement can be the object of another statement, which allows one to create directed labeled graphs.

RDF datasets "extend" the RDF data model with the notion of (named) graphs. In an RDF dataset, there is at most one unnamed graph and zero or more named graphs. Each graph contains a set of triples. The unnamed graph, typically called the default graph, has no identifier. Named graphs, on the other hand, are identified by a IRI. Triples in a named graph are thus four-tuples, as there is an additional component. RDF datasets thus provide a convenient way to organize knowledge, which we availed of in this project. For example, one can store information about people and their names from a document in one graph and information about marriages from marriage certificates in another.

---

[6]An Internationalized Resource Identifier (IRI) allows one to identify resources with a string of characters. The reader might be more familiar with Uniform Resource Identifiers (URIs). The IRI standard extended URIs by also including international characters (e.g., Korean). For this article, it suffices to say that IRIs allow one to identify and refer to resources on the Web with the HTTP and HTTPS protocols.
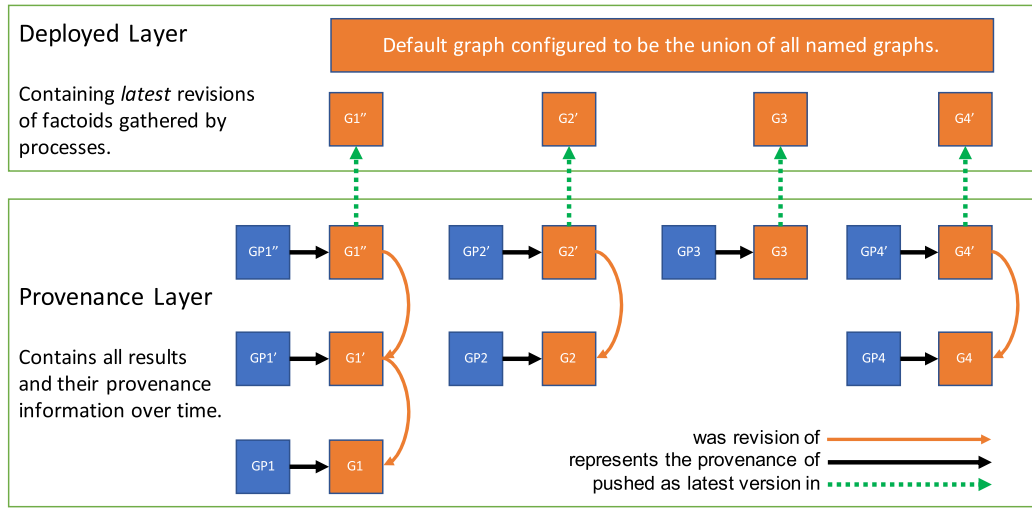
Fig. 3. Knowledge organization in the Beyond 2022 project. Blue squares represent named graphs containing provenance information about triples that have been generated via an activity (automatic extraction, transformation, manual entry, etc.). The triples that have been created are collected in separate named graphs (orange boxes). Only the latest versions of triples are deployed to the deployment graph.

3.2 The Knowledge Graph

For the KG, we decided to provide two endpoints.[7] Each endpoint provides a set of named graphs. The first endpoint is provided as the default gateway for information retrieval. The second provides a means to inspect the provenance of information and the evolution of the KG. Provenance information provides insights on a (RDF) resource's origin, such as who created that resource, when it was modified, or how it was created [38]. We visualize our approach in Figure 3. The first endpoint only provides the latest version of sets of factoids that have been created as the result of an activity (e.g., manual entry in spreadsheets transformed via mappings or automatic extraction). This is called the deployed layer. The factoids are grouped in named graphs—one per activity (orange boxes). This facilitates querying information on the whole graph. The provenance layer provides both the sets of factoids resulting from the activities above, as well as the provenance information (blue boxes) describing how the factoids came to be, as well as which factoids they replaced. For example, in Figure 3, the named graph GP1" provides information about the factoids in named graph G1", and GP1" also contains the statement that G1" is a revision of G1' (orange arrow).

Since provenance information and the history of sets of factoids are interested in specific scenarios and we assume that most users would be interested in the most recent available knowledge, we "push" the latest versions of sets of factoids to the deployed graph. As named graphs make the formulation of queries more verbose, we also decided to facilitate query formulation by servicing the union of all triples as the default graph of the deployment graph.

With the latest revisions of our KG's contents being separated into different named graphs, we can enforce work practices that ensure that each graph breaks down into a logical grouping of information. For example, each named graph G1, G2, and G3 may correspond to information gathered from three separate sources (e.g., books and manuscripts). These sources may conflict with each other in their description of entities. Yet, our method

---

[7]https://sparql.virtualtreasury.ie/b2022/sparql and https://sparql.virtualtreasury.ie/b2022-provenance/sparql.
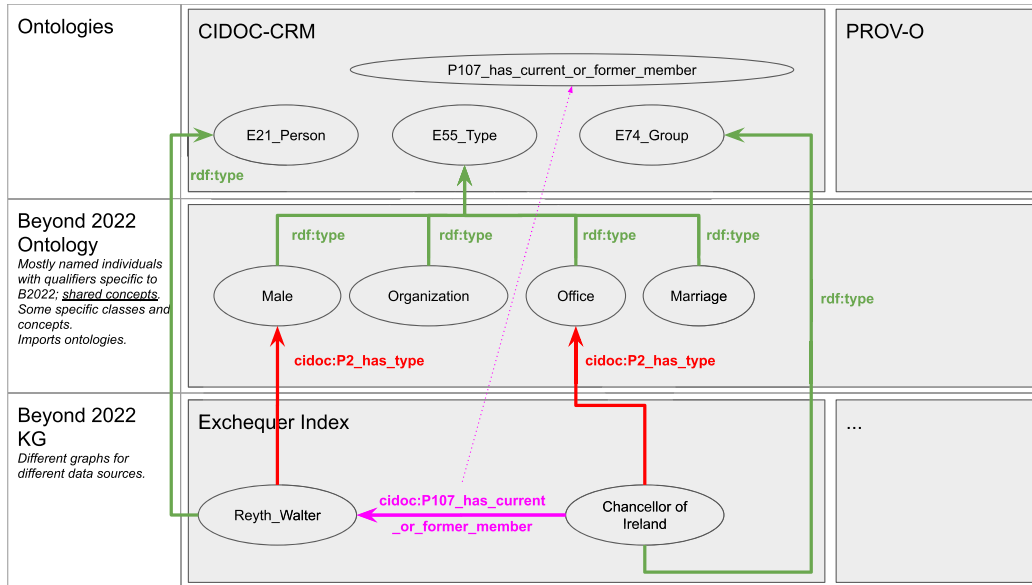
Fig. 4. The ontological layers in the Beyond 2022 KG. This figure illustrates that the Beyond 2022 ontology adopts and extends existing (standardized) ontologies, mostly to introduce concepts specific to the project, and then uses these to populate the KG.

of structuring information means we can keep these potential conflicts separate until we can agree on how to resolve them and ultimately ingest them into the deployed layer.

Additionally, triples that are produced by an NLP pipeline can be kept in their own respective named graphs enabling us to easily accept or disregard the information provided by an automatic process. While NLP techniques provide a convenient approach to enriching the KG, they are not as authoritative as the carefully curated factoids of a historian. Even when a historian and an NLP pipeline process the same corpus, we store the results of an NLP pipeline in a different named graph. If a scholar does not sufficiently trust triples gathered with AI techniques, for instance, then the scholar can request not to consider those named graphs whilst querying the graph. In summary, named graphs are used to capture the different interpretations (manual or "automatic" if we can call it that for automatic approaches) of different sources.

While the ontology, which we will discuss in the next section, is made available as a separate artifact according to Linked Data principles, we also store the ontology and its provenance information in the triplestore. While it is true that the ontology is "stored" in two different locations, this approach makes the formulation of SPARQL queries over the types (ontology) and entities (instances) much easier.

4   ONTOLOGY ENGINEERING IN BEYOND 2022

We will describe the ontology of the Beyond 2022 KG and its construction in this section. The key concepts were identified via competency questions. We also adopt and extend an ontology widely adopted in the cultural heritage domain. This ontology, while generic, needs to be extended for us to describe concepts that are specific to the Beyond 2022 project. This means that there are two ontological layers in the Beyond 2022 KG, as seen from Figure 4, which are used to populate the KG. The concepts introduced in our ontology needed to be defined and formalized. We will thus also describe these processes in this section.

The Beyond 2022 ontology, which primarily consists of named individuals for qualifying entities in CIDOC-CRM, has been published according to best practices in Linked Data. Its documentation and behavior according to Linked Data principles have been generated with WIDOCO [10].

4.1 Competency Questions

Competency questions [14] allow a group of stakeholders to formulate a system's requirements in terms of the questions that such a system should (help) answer. Any system that can answer these questions is deemed adequate, no matter how the information is organized. In other words, differently structured ontologies may provide answers to the same set of competency questions.

The competency questions were obtained from subject matter experts in the Beyond 2022 consortium. We surveyed a team of four historians involved with the project and asked them to submit several questions which they would expect the KG would be able to answer. Instructions were kept deliberately vague. The intention was to encourage the historians to express their highest expectations for what the KG would be able to do. It was felt that this was necessary as many of the historians had not worked with KGs before. Giving them a large degree of freedom helped the computer science team to not only design the structure of the KG, but also perceive how historians viewed it as a research tool.

The responses comprised 54 competency questions that primarily focused on entities such as **people**, **organizations**, and **places**, although there were some interesting outliers pertaining to trade. Another type of entity that emerged as being of importance is the notion of **office**, positions held by people (e.g., "Who was the [office name] in [year]?"). Those competency questions were then refined and, where appropriate, grouped into a list. That list was then sent around to the consortium with the question to rank those questions in terms of perceived importance for the system to answer. The 10 highest-ranked competency questions are:

(1) Show me birth|death|marriage records in [place] in [year].
(2) Is [name] an Irish name?
(3) Were there any [surname] living near the [surname]?
(4) What was the average rent per acre in [place name] in [century]?
(5) Did [family] have more land than [family]?
(6) What was the average rent per acre in [place name] in the [century]?
(7) Were there any [surname] within 20 miles of [place name] just before the famine?
(8) Show me [religion] as a percentage of the population in [year] as a pie chart.
(9) What was the average price of land per acre in [century] [county]?
(10) Show me maps of [place name].

Many of these questions relied on concepts currently outside the historical period we chose to process, namely historical figures mentioned in [2], which deals with 13th and 14th century Ireland. Examples of such concepts pertain to census data and vital records, which appeared much later in the history of Ireland.

Competency questions are the questions that the KG should answer or *help* answer. In the case of the pie chart mentioned by one of the competency questions, an application on top of the KG will generate that chart upon receiving the results of such a query.

While the Beyond 2022 KG cannot cater to all of the consortium members' projects and needs, we identified several commonalities that fit within Beyond 2022's overarching goal of providing information on entities mentioned in the digitized assets.

4.2 Adopting and Extending PROV-O and CIDOC-CRM

Provenance information provides insights on a resource's origin, such as who created that resource, when it was modified or how it was created [38]. Provenance, as stated in [16], is key in evaluating the quality of, and establishing trust in information on the Web. In our approach, provenance information is represented with
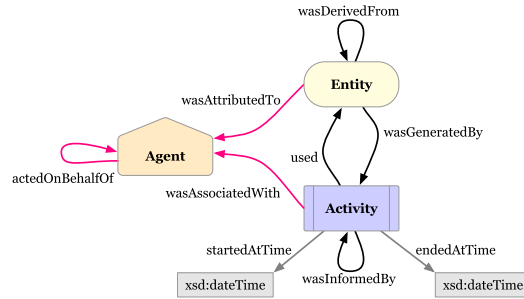
Fig. 5. Core concepts and relations in PROV-O from [23], Copyright (c) 2011-2013 W3C(r) (MIT, ERCIM, Keio, Beihang).

PROV-O [23], which is a W3C Recommendation. PROV-O provides a way to capture this information in (specializations) of the classes entity, activity, and agent, and their relations. PROV-O's core concepts and relations (shown in Figure 5) provide a good starting point for describing the activities, artifacts, and software used to generate the KG. Within Beyond 2022, we only had to specialize these concepts. This exercise is relatively straightforward; e.g., the introduction of a transformation activity to represent the transformation of spreadsheets (saved as CSV files) into RDF (see Section 6).

More important is the adoption and extension of the domain model provided by CIDOC-CRM.[8] CIDOC-CRM provides a sufficiently abstract model for capturing most of the concepts and relations in the domain. CIDOC-CRM provides a broad set of classes and relations for representing cultural sector data. For our KG, we adopted the OWL implementation of the CIDOC-CRM [11].

In [24], the authors identified some issues related to scalability and inferencing with this model. Issues include that the model lacks "common" concepts and relations such as marriage, gender,... requiring those that avail of CIDOC-CRM to declare those concepts and relations, especially when "importing" extension would lead to importing other irrelevant and potentially conflicting knowledge. The concept of marriage and gender may have been declared in an extension of CIDOC-CRM for social sciences, for example, but importing that extension into our own ontology will require that we import *all* of the extension's axioms and may lead to undesirable results when engaging with the KG.

Other issues pertain to the fuzziness of, e.g., time intervals for which no adequate inference engine is part of the Semantic Web stack. As we are not concerned with reasoning over the model and CIDOC-CRM is an established metamodel in the cultural heritage domain, we still choose to avail of it. Within the project, however, we aimed at minimizing the need to extend the ontology. Whenever CIDOC-CRM could not provide types or relations that fitted our needs, of which we will provide examples below, we favored introducing bespoke types that are related to entities via the `cidoc:P2_has_type` predicate (instead of creating instances with `rdf:type`). The goal was to ensure that our KG has maximum backward compatibility with the CIDOC-CRM XML specification. The lack of common types and the extensive use of CIDOC-CRM's typing system renders CIDOC-CRM quite complex. Its complexity will also be discussed in Section 8.

Examples of bespoke types (all of which are instances of `cidoc:E55_Type`) we introduced are[9]: `:Floruit` to represent the period in which a person "flourished" i.e., we are aware they were active during this period; `:Name` for qualifying the normalized appellations that historians curated and `:NameVariant` for qualifying the variant

appellations[10]; :Office, :Occupation, and :Rank for qualifying some of the professional groups a person belonged to. As the factoids created by historians came from sources (called Authority Documents in CIDOC-CRM), we also introduced instances of cidoc:E55_Type for the different types of resources to facilitate information retrieval. Particular attention was given to names. As historians will often start from a person's name, a person may have many different variants in spelling (even across languages), and a name can be used to identify many people, it is useful to consider names as first-class citizens in the KG–an aspect recognized by CIDOC-CRM.

Concepts that were introduced in the B2022 ontology emerged both in a top-down fashion (using ontology engineering methodologies), and in a bottom-up fashion (by looking at concepts that were introduced by historians). We will describe both approaches in Sections 4.3 and 4.4.

4.3 Top-down Ontology Engineering with DOGMA

A particular challenge in the creation of Beyond 2022's ontology is facilitating stakeholder participation: i.e., enabling the historians in modeling their UoD. While it is not expected that historians become sufficiently proficient with the formalisms and models underpinning the KG, the knowledge engineers in the project needed to rely on their domain expertise to ensure that graph is meaningful. To this end, we availed of a particular knowledge engineering methodology called DOGMA [20], which allows one to represent a UoD using their natural language.

In DOGMA, people model using *lexons*, which are quintuples consisting of a context label, term, role, co-role, and co-term. Each lexon represents a relationship between two concepts that hold in a particular context. The relationships are then broken down into the role and co-role. While the relationship has no direction, the roles have to be read in a particular way. For instance, the lexon <Beyond 2022, Castle, has, part of, Wall> states that in the context of Beyond 2022, castles have walls and walls are part of a castle.[11]

We do not rely on a bespoke tool, instead we ask historians to create lexons in a spreadsheet (see Figure 6). This spreadsheet is shared with the knowledge engineers in the project. Knowledge engineers analyze the lexons and interact with the historians until they converge to a mutual understanding. Knowledge engineers also assist the historians in the modeling exercise. In Figure 6, one can see that an additional column was added to represent whether and how a lexon is included in the ontology. As we avail of CIDOC-CRM, some of the lexons were straightforward to implement, while others required certain design patterns. For the taxonomic relationships, we avail of SKOS [34]. The rdfs:subClassOf predicate to relate more specific types to more generic types is reflexive, which means that each concept is also a subclass of itself. The properties of rdfs:subClassOf did not align with Beyond 2022's requirements and we therefore adopted SKOS where the skos:broader property is not reflexive. Some examples, which are used to illustrate the approach to historians, are shown in the figure.

Terms might evoke a specific concept amongst historians, but we are not assured whether every historian has the same concept in mind, lest agree on the meaning of the concept. We gather definitions (also called glosses) from the historians in a separate spreadsheet. It is pretty straightforward for the ontology engineers to integrate these definitions into the ontology. Currently, the consortium uses color-coding to keep track of changes (e.g., green to indicate that the definition has been put in the ontology, and yellow to indicate changes made by historians).

4.4 Bottom-up Emergence of Concepts

Put simply, the ontology captures the types of entities. In other words, they should be sets or categories of things. One should see the concept of "Church", but not a specific church (e.g., "St. Mary's in Dublin"). The

---

[10]Normalized appellations and their variants are related with the property cidoc:P139_has_alternative_form. We have chosen to create types for the two (and both have different IRI strategies) so that one can easily look up and filter either and more easily distinguish between the two.

[11]DOGMA furthermore allows one to declare constraints using lexons and a controlled natural language. One can, for instance, state: EACH Castle has AT LEAST 1 Wall. In Beyond 2022, we do not yet avail of such constraints as the ontology is meant to support interoperability and foresee many uses. As [33] observed, the reusability of a schema and ontology goes down as the number of constraints increases.

Context	Term	role	co-role	co-Term	Implementation Details		
Example	Author	writes	written by	Publication			
The following method for ontology elicitation is based on the DOGMA approach, which uses a controlled natural language to structure a universe of discourse.							
In this approach, you structure the world using binary relations, and each relation consists of a role and a co-role. That should result in two meaningful stences read in the following way:							
IN THE CONTEXT OF [CONTEXT] EACH [TERM] [ROLE] [CO-TERM], and EACH [CO-TERM] [CO-ROLE] [TERM].							
In the example above, we thus have: "In the context of "Example", EACH Author writes [a] Publication, and EACH Publication [is] written by [an] Author.							
Jarrar, M., & Meersman, R. (2002, October). Formal ontology engineering in the dogma approach. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 1238-1254). Springer, Berlin, Heidelberg.							
Why the role and co-role? We as humans, we tend to use the same roles for different types of relations. Thinking about the other direction "forces" one to think about the true meaning of the relationship. For instance, a person is subject to a king, and a person is subject to a law. One way to disambiguate those different "subject to" relations is to: 1) think of the co-role, and then 2) refine the names when necessary. Remember, humans are able to disambiguate from context and computer have a harder time. This is important for the ontology engineer if it is key for the competency questions.							
Below are some examples which have been implemented in the Beyond 2022 ontology.							
taxonomy	Civil Organisation	is a	subsumes	Organisation	skos:broader		
taxonomy	City	is a	type of	Civil Administrative Unit (Municipal)	skos:broader		
buildings	Castle	has	part-of	Wall	cidoc:P89_falls_within (between instances)		
organisations	Organisation	reports to	reported to by	Organisation	propositional objects		
Use this document to formalize aspects of the universe of discourse that have either not been modelled, or need to be refined. The Ontology Engineer can use the comments section to indicate whether and how they are implemented. All the examples above are inspired by Beyond 2022, but not necessarily implemented as such–they are merely examples.							

Fig. 6. Spreadsheet handed over to historians for capturing lexons. Note that the sixth column is used to indicate whether a lexon has been implemented, and how. The how depends on the ontologies we have adopted as the foundation of the Beyond 2022 ontology. This spreadsheet is part of a suite of spreadsheets that historians use to populate the Beyond 2022 KG.

creation of entities, which is discussed in Section 5 is a different challenge. However, we do not want historians to be constrained to the concepts available in the ontology. We therefore allow historians to create concepts "on the fly" when creating entities. This is part of our approach to bottom-up elicitation of concepts.

When historians introduce a label in a spreadsheet that refers to a concept, that label is used by the mapping to create an IRI for that concept. As we model our KG as per CIDOC-CRM, these result in instanced of cidoc:E55_Type that are connected with an instance via cidoc:P2_has_type. If the IRI exists in the ontology, the connection is with the ontology is "automatically" made. Concepts provided by historians that do not appear in the ontology in this manner are easily retrieved with SPARQL. In Section 3, we described how information from different agents are stored in different named graphs. It is thus only a matter of looking for all types that do not appear in the ontology with the SPARQL MINUS operator, as shown below. The first part of this SPARQL query retrieves all the types used in the KG. The second part removes, from that list, the types that have been declared in the ontology with the MINUS operator.

```
PREFIX cidoc: <http://erlangen-crm.org/current/>
SELECT DISTINCT ?type WHERE {
  # All the types in used in the knowledge graph
  { GRAPH ?g { ?thing cidoc:P2_has_type ?type} }
  MINUS
  # All the types declared in the ontology
  { GRAPH <http://ont.virtualtreasury.ie/ontology> { ?type a cidoc:E55_Type  } }
}
```

5 FACTOID ELICITATION WITH SPREADSHEETS

The uptake of RDF by memory institutions is known to be problematic, as evidenced by surveys conducted in [26]. While there are quite a few noteworthy projects, issues with Semantic Web technologies include tooling
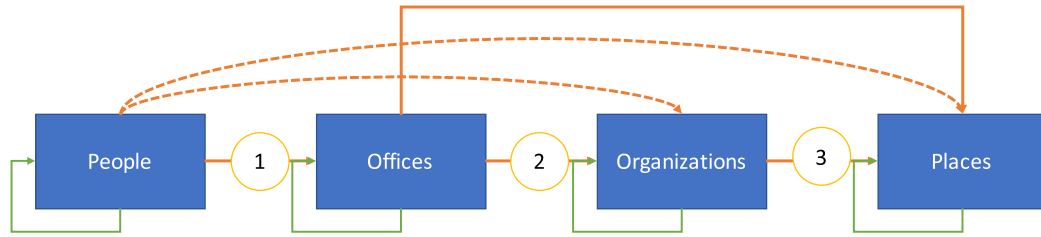
Fig. 7. The relationships between the core spreadsheets for factoid gathering. Green arrows indicate specific relationships between entities of the same type. Orange arrows indicate relationships between entities of different types. When the arrow is full, they denote specific relationships. When the arrow is dashed, one can declare *a* relationship without providing the type of that relationship.

and some level of background knowledge that is required to understand and even appreciate some of its quirks (e.g., concerning blank node identifiers and the open-world assumption).

One technology with which people are arguably more comfortable with are spreadsheets, which can be considered as non-normalized tabular data. For instance, tabular data stored as either CSV or TSV are an accessible and convenient way of storing and sharing data and is a practice adopted in many projects. We thus developed a suite of spreadsheets allowing historians to capture information about the entities identified in the project: people, offices, organizations, and places. Multiple spreadsheets were developed, though we will first elaborate on the first four—one for each of the concepts. The relationships between these four sheets are illustrated in Figure 7. While the sheets are independent, the editorial guidelines allow historians to follow a process to gather information about these concepts.

In the People-sheet (see Figure 8 for an example), historians gather information about people and their interrelationships (represented by the reflexive arrow). This sheet also captures information about a person's tenure, rank, status, and so on. The information about a person's tenure is captured as a relationship with an office (zero or more). In the KG, this results in an instance of `cidoc:E21_Person` being related to an instance of `cidoc:E74_Group` of the `cidoc:P2_has_type :Office`. If a historian wishes to provide more information about that office, then the historian can represent this in the second sheet.

In the Offices-sheet, historians are able to represent information about offices and their interrelationships (e.g., offices overseeing other offices). This sheet also captures the relationships between offices and organizations, and offices and places. When information about organizations or places is entered, these can be further "completed" in their respective sheets. Both Organization and Places-sheets capture, much like the Offices-sheet, the information about these entities and their interrelationships.

Note that historians are not obliged to complete all sheets. The information about an office entered in the Person-sheet, for instance, will generate an instance of that office, its appellation, and its relationship with that person. The suite of sheets is meant to allow historians to represent in sufficiently granular detail the entities appearing in sources.

Through workshops with a team of historians, we also recognized the need for stating *a* (and not *a specific*) relationship between a person and an organization or place. Providing support for arbitrary relationships in the spreadsheets would not scale well (e.g., separate columns for each type of relationship or key-value pairs in spreadsheets) would become difficult to fill in and manage. Historians, however, just expressed a desire for a way to declare superficial relationships between people and other entities in the first sheet. We thus have provided support to relate people with zero or more organizations and places in the Person-sheet. As CIDOC-CRM does not have a property for such superficial relations (e.g., "related with"), we had to avail of instances of `cidoc:E89_Propositional_Object`. A `cidoc:E89_Propositional_Object` is a reified relationship, which means that an instance of `cidoc:E89_Propositional_Object` represents that relationships and will have

Fig. 8. Excerpt from a spreadsheet currently curated by a historian. In this spreadsheet, information about various people mentioned by the *Irish Exchequer Payments, 1270-1446* [2] (see Section 7) is captured. Information includes, when available, alternative spellings of names, their offices and ranks, and their dates of birth and death. Some records also contain information on spouses and children. We provide a sample of this spreadsheet in [6].

properties to both the person and the place or organization. So when a historian declares that "John" is related to the location "Meath", that will result in a propositional object referring to both "John" and "Meath" with `cidoc:P67_refers_to`. More detailed information on the relationships between people and other entities can be provided by filling all the spreadsheets.

These four sheets capture the necessary information to relate the four entities. When more detailed information about a particular relationship is required, such as the time period, then these four sheets do not scale well. We therefore developed additional spreadsheets: one for gathering information about relationships between places and organizations (e.g., the parliament moved to different locations throughout history), and one for people and offices (e.g., to capture that a person was appointed to an office four times). Within the project, emphasis is initially put on the core sheets as this allows us to capture the Chief Governors of Ireland, for instance. Detailed relationships between entities (e.g., the time period(s) between a person and an office they held) were "nice to have."

Whilst entering data, historians have the ability to either enter: (1) strings for the declaration of an entity. If the same values within and across sheets are used for the same person, then this results in the same IRIs in the KG, (2) internal IDs that hold within sheets, e.g., to relate spouses, and (3) HTTP and HTTPS IRIs, which is useful when entities already exist in the Beyond 2022 KG. SPARQL Faceter [22] is one such tool, allowing one to look for an entity's IRI based on faceted search.

It is important that historians do not delete any of the columns, but they are free to rearrange and even hide columns to tailor the sheets to their needs and workflows. The RDF that will be generated from these sheets will be stored into different named graphs as to separate the different factoids and interpretations, but we want to support the provenance of these interpretations. Each sheet therefore provides various columns to accommodate those:

(1) Source and index numbers to indicate where the entity was mentioned in the "primary" resource;
(2) Several reference columns to allow the historians to reference additional material consulted to complete the information about an entity;
(3) For people, we also provide columns for entries in the Oxford Dictionary of National Biography and the Dictionary of Irish Biography;
(4) A comment column where information about the factoids can be provided (e.g., to cope with uncertainty with dates).

The support for capturing the "uncertainty" for each factoid was considered, but this would have not been scalable. We therefore provided one column, which will result in a note (`cidoc:P3_has_note`) for the entity in that particular row.
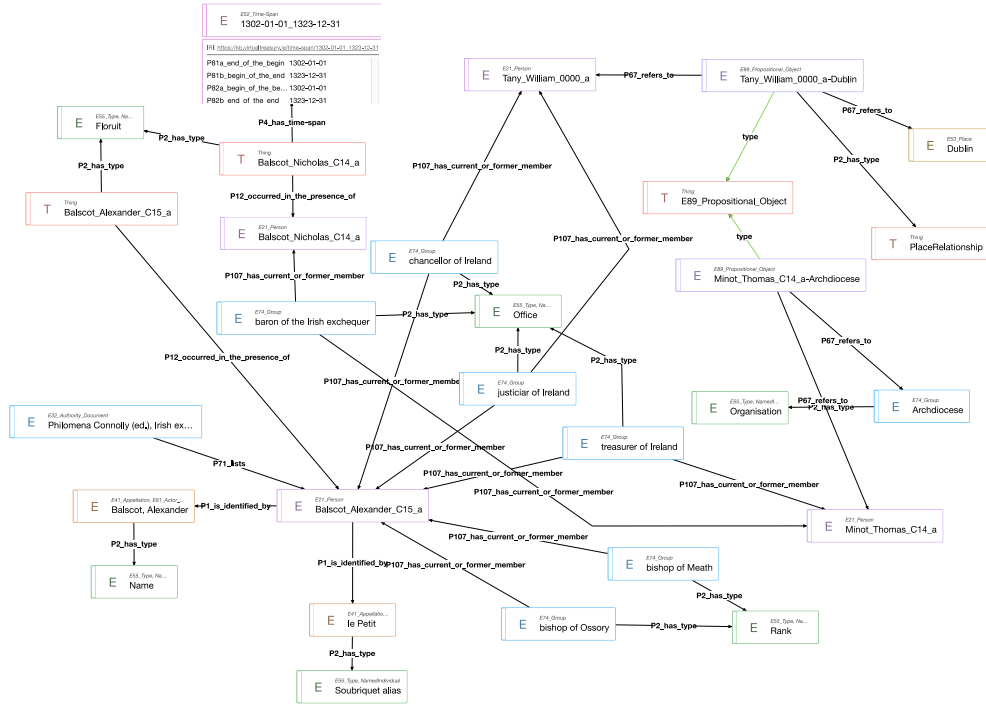
Fig. 9. Excerpt from the Beyond 2022 KG using the triples coming from the spreadsheets in Figure 8. This image, created with Ontodia, provides examples of our core concepts, namely: people, offices, organizations, and places. One can zoom in on this vector-based image to see all the details.

The editorial guidelines on how to fill in those spreadsheets are crucial to managing the identification of entities, which result in their IRIs. Within the Beyond 2022 project, historians have consulted with knowledge engineers on an identification strategy (i.e., "Which attributes are necessary to identify instances, and how"). The historians then took the lead in formulating these editorial guidelines. It is up to a historian to look for the prior existence of an entity. If no such entity is found, the one entering the data must follow these guidelines. The guidelines provide instructions on, among others: how to format values (names, dates, discriminators, etc.), capitalization, and which name should be chosen as the preferred identifier. This means that historians may need to consult additional sources to find the preferred identifier (e.g., when a person is referred to by a sobriquet).

5.1 Generating RDF from the Spreadsheets

We already alluded to generating RDF from the information contained in the spreadsheets. In our approach, the spreadsheets are stored as CSV files and transformed into RDF via R2RML [4], a W3C Recommendation to transform relational data into RDF via a set of mappings. We avail of [7], which allows us to access the CSV files as relational databases. Those mappings prescribe how the data contained in those spreadsheets should be transformed into entities and relationships according to CIDOC-CRM and our ontology. The use of R2RML thus allows for a scalable and declarative ingestion pipeline. The base IRI of the KG's entities is https://kb.virtualtreasury.ie/. Figure 9 provides a figure containing some of the instances, types, and interrelationships of the Beyond 2022 KG. This image was created with Ontodia [28], an application to explore KGs visually. This figure provides examples of the main concepts: people, offices, organizations, and places.

As spreadsheets are not a substitution for relational databases and subject matter experts have some freedom in defining and using the spreadsheet "schema", we had to account for non-normalized columns; columns in which cells could contain more than one value. Either one has access to database functions that transform non-normalized values into a relation, which are not part of the SQL 2008 standard, or one either needs to avail of recursive queries in the mappings. Fortunately, schemas developed by historians are to be reused (and may be extended), and hence mappings needed to be written once by the knowledge engineers in the project.

Finally, we note that both spreadsheets and their corresponding mappings are easy to extend. While the goal of the Beyond 2022 KG is to represent information relevant to the project using common concepts, it is possible for historians to add additional columns. The addition of bespoke columns in the spreadsheets does not hamper the RDF generation process. The mappings for these columns, if relevant for the project, can subsequently be engineered.

6   CREATING AND MANAGING THE KNOWLEDGE GRAPH

The knowledge platform of the Beyond 2022 project supports information extracted from various sources via different means: entity recognition in documents and manually curated data are two of them. We focus on the manual curation of the KG. Historians, who are familiar with spreadsheets, capture information according to a particular schema. Those files are then ingested into the triplestore via a tool only accessible to them (see Figure 10).

Using this tool, historians can upload a CSV file, choose a mapping and, optionally, choose the named graph they want to override. The latter is useful when a historian wants to amend or add to their already deployed work. One may notice that there are three stages. Upon uploading a CSV file, the user first needs to transform the CSV into RDF and check the output with a set of SHACL [21] constraints.[12] That will allow the historian to assess whether the data contained in the CSV is fit for deployment. Before one can deploy the resulting RDF, however, one first needs to check whether the union of the resulting RDF and the other named graphs (minus the graph to be replaced if applicable) does not violate the SHACL constraints. The SHACL constraints are pretty straightforward, they check whether:

— The date of birth is lower than the date of death;
— The dates related to one's floruit falls within one's date of birth and death;
— One has at most one gender;
— ...

When all is in order according to the SHACL contraints,[13] the historian can deploy the new graph. This process can be summarized as follows:

— Two IRIs, $u1$ and $u2$, are generated. One for the named graph that will contain the generated triples and one for the provenance graph. The suffix of both $u1$ and $u2$ are the same, their IRIs differ by one component

---

[12]Creating SHACL constraints was not always as intuitive. One would assume that one could compare a person's birth date with that same person's date of death starting from a person. The "starting point" in SHACL is called a focus node. In SHACL, one can only compare, starting from a focus node, one complex path (e.g., the date of birth via the event connected to a person–multiple "hops") and a simple path (only one "hop"). This meant that we had to model this constraint starting from the date of birth (to have the date as the immediate value) and then retrieve the other date via person and its date of death. While possible, this made modeling constraints quite convoluted. We, therefore, modeled constraints as SPARQL-validators. The semantics of the rule is "hidden" in the SPARQL query, but the constraints are, in our opinion, easier to read and manage.

[13]SHACL validates the resulting graphs, and not the CSV file. There may be instances in which there are errors in the CSV file that are not picked on by SHACL. One such case is when a record does not contain enough information to generate RDF. We can validate CSV files with CSVW-validator (https://github.com/malyvoj3/csvw-validator.). This tool allows us to use the CSV-W [35] language to declare constraints on spreadsheets. This tool has not yet been integrated into the prototype, however, we assume that the CSV files have been checked.

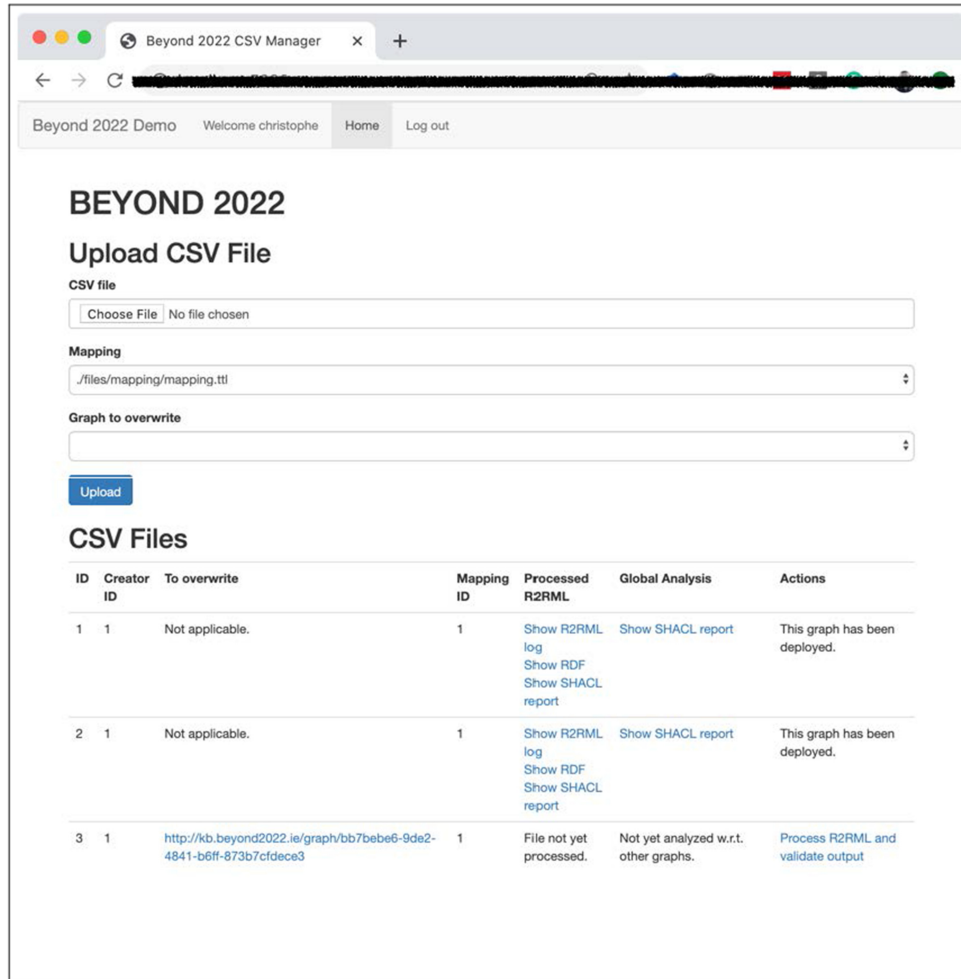[14]https://github.com/malyvoj3/csvw-validator.

Fig. 10. Managing the CSV files, uploaded by historians, which are transformed into RDF and deployed the named graphs.

in the path. The IRI $u1$ will contain /graph/ and $u2$ will contain /provenance-graph/. If no labels are
provided for the IRIs, a UUID for both IRIs will be generated.
— A provenance graph is generated for the transformation activity. This provenance graph will refer to the
named graph $u2$. If the named graph $u2$ is used to revise another graph $u3$, then the provenance graph will
contain the statement: `<u2> prov:wasRevisionOf <u3>`.
— In the "Deployed Graph", the graph to be revised (if applicable) is removed.
— The generated triples are loaded into graph $u1$ on both the deployed and provenance layer.
— The provenance graph is loaded into graph $u2$ on the provenance layer.

The following RDF TURTLE snippet contains triples from the provenance graph. We have shortened the file
paths in the snipped. These triples, stored in `provenance-graph/places-v2` contain information on a mapping
activity. Here, the mapping activity, associated with a member of the team, generated a revision of the triples
stored in `graph/places-v1`, which are stored in `graph/places-v2`.

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<https://kb.virtualtreasury.ie/graph/places-v2>
  a prov:Entity ;
  prov:wasGeneratedBy  [
    a prov:Activity , <https://ont.virtualtreasury.ie/ontology#R2RML_Mapping_Activity> ;
    rdfs:label "Mapping Activity" ;
    prov:endedAtTime "2020-02-12T14:36:04.04"^^xsd:dateTime ;
    prov:startedAtTime "2020-02-12T14:36:02.778"^^xsd:dateTime ;
    prov:used <file:///.../SAMPLEENTITYLISTPC.csv> , <file:///.../mapping.ttl> ;
    prov:wasAssociatedWith <http://virtualtreasury.ie/team#christophe>
  ] ;
  prov:wasRevisionOf <https://kb.virtualtreasury.ie/graph/places-v1> ;
.
<file:///.../mapping.ttl> a prov:Entity , <https://ont.virtualtreasury.ie/ontology#R2RML_Mapping> .
<http://virtualtreasury.ie/team#christophe> a prov:Agent .
<file:///.../SAMPLEENTITYLISTPC.csv> a prov:Entity , <https://ont.virtualtreasury.ie/ontology#CSV_File> .
```

7 CASE STUDY: THE IRISH EXCHEQUER INDEX

The initial development of Beyond 2022's KG depended upon the use of a test case consisting of robust, clean historical data which could be used to develop both spreadsheet schemas and a bespoke ontology for Beyond 2022. Both the spreadsheet schemas and the ontology needed to accommodate as broad a range of historical scenarios or instances as possible, encompassing not only a broad range of themes or topics, but also a chronological span of over seven hundred years. A primary objective of this project is to ensure that Beyond 2022's KG will "unlock" —or enable meaningful exploration of—Beyond 2022's reconstructed archival collections, and identify links between these collections and those held in other national or international institutions.

The initial test case used for the KG was designed to interlink with another central strand of Beyond 2022: the Medieval Exchequer Gold Seam. This strand of the project aims to reconstruct entire series of financial records drawn from the financial powerhouse of English government in Ireland, the medieval Irish exchequer, records of which date back to the twelfth century and, indeed, to the beginnings of Beyond 2022's reconstituted archive. The Exchequer Gold Seam involves the transcription and translation of many hundreds of medieval records now held in The National Archives (UK) which have, until the inception of this project, never been reconstructed in their entirety. As such, the data used for the KG's initial test case was drawn from a related collection of immense importance to the history of the medieval Irish exchequer, and indeed to the history of Ireland in the Middle Ages. In 1998, Philomena Connolly's calendar of *Irish Exchequer Payments, 1270-1446* [2] was published by the Irish Manuscripts Commission. Connolly was Ireland's foremost authority on the medieval Irish exchequer, and her meticulously researched and translated calendar of the medieval exchequer's issue rolls provided a model dataset for the KG test case. The index provided in this text records over 2,000 historical individuals, all of whom have now been included as person entities within the KG. Many of these person-entities will appear frequently in the historical collections processed by Beyond 2022, and also further afield.[15]

The records of the medieval Irish exchequer of Ireland lend themselves very well to such a project, chiefly because they contain precisely the type of historical data (and indeed, historical "problems") required to design a KG which can accommodate the vast and varied historical collections which will be processed by Beyond 2022. The records of the medieval exchequer are rich in terms of person, place, and office entities—the types of entities which have been identified by subject matter experts as key to understanding fundamental developments within Ireland's history from the twelfth to the twentieth century. The activities and reach of the Irish Exchequer touched upon nearly all aspects of political, social, and economical life in the lordship of Ireland in the Middle Ages; its existence continued, albeit in a different form, in early modern and modern Ireland too. The records

---

[15]A sample of the data has been made available in [6].

Table 1. Overview of the Various Named Graphs (From Sources)
Currently Deployed in the Beyond 2022 KG

Description of named graph	# of triples
Irish Exchequer Index based on [2]	
People	76,345
Additional triples on offices	1,735
Additional triples on places	189
Additional triples on organizations (ongoing work)	170
Movement of organizations over time (ongoing work)	81
Other sources	
Modern townlands from OSi and OSNI	1,321,008
Chief Governors of Ireland based on [27] (ongoing)	76,345
Down Survey (historical townlands) (ongoing)	11,684
Regnal Years	6,454
Creators of artifacts in the PROI	2,446
The Beyond 2022 ontology	804

of Ireland's medieval exchequer therefore shed light upon not only records of high-ranking individuals—the lieutenants, justiciars, chancellors, and treasurers of Ireland, all of whom headed central institutions of Ireland's medieval administration under the English crown - but also, and perhaps more importantly, upon the records of ordinary people. For the development of the Beyond 2022's KG, then, the value of this source material lies in its potential to uncover and reveal through KG technologies a vast range of governmental and societal relationships, structures and hierarchies in a way that allows historians to conduct meaningful research, and identify connections which are not otherwise apparent, or would require extensive time and research to reveal (as shown, for example, in Figure 11). Equally important to the development of the Beyond 2022 ontology, however, is the latent potential also contained within the records of the Irish exchequer. Their importance is not limited to their capacity to reveal the intricacies of Ireland's medieval government and life. Rather, the objective in using this data was to investigate, discover, reveal, and reconcile historical problems and scenarios which spanned beyond the medieval: that is, historical scenarios, relationships, connections, peculiarities, concepts—even ideologies—that occurred across broad swathes of time and place. Connolly's work has enabled us to extract not only person-entities, but also offices associated with individuals mentioned within these records and, consequently, medieval governmental organizations and indeed the relationships between both individual offices and their larger organizations. By distilling from this process the most fundamental ordering concepts of government and society, and by categorizing them using broad, abstract concepts, Beyond 2022's knowledge engineers and historians have been able to establish a set of hierarchies or templates which can be used to map Ireland's administrative, governmental, and societal structures across a chronological span of just over seven centuries.

At the time of writing, the deployment graph (i.e., the latest versions of named graphs containing factoids) of Beyond 2022 KG contains 1,497,261 triples. In Table 1, we provide an overview of all the triples in each named graph in the deployment graph. We also indicate which parts of the name graph are still being completed or have not been considered as completed. Of those triples, 78,520 belong to the Irish Exchequer Index as a source. Most of the triples come from structured sources on geographic features and their boundaries. It seems that the KG is skewed towards geographical data, but that is merely the result of transforming the existing datasets on thousands of townlands.

Table 2 provides an overview of the various types (both via `rdf:type` and `cidoc:P2_has_type`) and the number of instances of that type after transforming the person spreadsheet created for the Irish Exchequer Index. This spreadsheet contains information on 2,201 people. There are 12 instances of `cidoc:E21_Person` that

Table 2. Number of Instances Per Type After Generating RDF from the First Spreadsheet of the
Irish Exchequer Index

Type	Count	Type	Count
cidoc:E41_Appellation	7,367	b2022:Floruit	308
cidoc:E81_Actor_Appellation	7,060	cidoc:E52_Time-Span	248
b2022:NameVariant	4,471	b2022:Office	246
cidoc:E21_Person	2,213	b2022:Forename	236
cidoc:E67_Birth	2,201	b2022:Occupation	226
cidoc:E69_Death	2,201	b2022:PatronymicsMatronymics	95
b2022:Male	2,149	b2022:Rank	81
cidoc:E89_Propositional_Object	1,442	cidoc:E53_Place	72
b2022:Name	1,438	b2022:SoubriquetAlias	55
b2022:Surname	765	b2022:ODNB	45
b2022:OrganisationRelationship	712	b2022:DIB	37
cidoc:E32_Authority_Document	706	b2022:Female	36
cidoc:E74_Group	584	b2022:Organisation	31
b2022:PlaceRelationship	422	b2022:Marriage	22
b2022:Career	308	cidoc:E85_Joining	22

are stored as blank nodes. Some people have been mentioned in passing and were judged not to be of particular interest (see Section 3).

7.1 Demonstration

The semantic technologies adopted in this project allow the team to adopt existing tooling that can process RDF. While the semantic web query language (SPARQL) itself might present too steep a learning curve for certain historians, tools such as Ontodia [28] and SPARQL Faceter [22] will allow historians to visually explore the KG.

Historians in Beyond 2022 have already seen how one can easily discover the people associated with certain offices (e.g., "chancellor of Ireland") and the overlap between different offices using Ontodia. In Figure 11, we demonstrate how one can discover people that have filled several positions, with one person being associated with all three; a certain Alexandre Balscot. We start by selecting the office "chancellor of Ireland" (Figure 11(1)). For the purpose of this article, we only place 8 of the 26 people that held this office in our KG on the diagram. We then select the office "treasurer of Ireland" (Figure 11(2)). This office has links to 46 people and we can already see that 4 of the initial 8 people held the two positions. Finally, we add the office of "justiciar of Ireland" (Figure 11(3)). This office has links to 40 people. We immediately see that three people on the diagram also held that office. Only one person on the diagram held the three offices. The IRI of that person can then be used to retrieve a page with additional information about that person.

This article's medium does not allow us to show the diagram with all 112 people elegantly.[16] Important is that the KG representation of the manuscripts in combination with tools like Ontodia allow subject matter experts (and other users, for that matter) to discover information that might have taken a lot more time going through manuscripts.

When one uses tools to explore the KG, be it via Ontodia by exploring the relationships of an entity or via SPARQL Faceter with facets mostly on labels and appellations, users can follow these IRIs to obtain

---

[16]For a video of this demonstration involving all individuals in the KG, we refer to https://www.youtube.com/watch?v=X3dVWkVOrS0&t= 3859s.
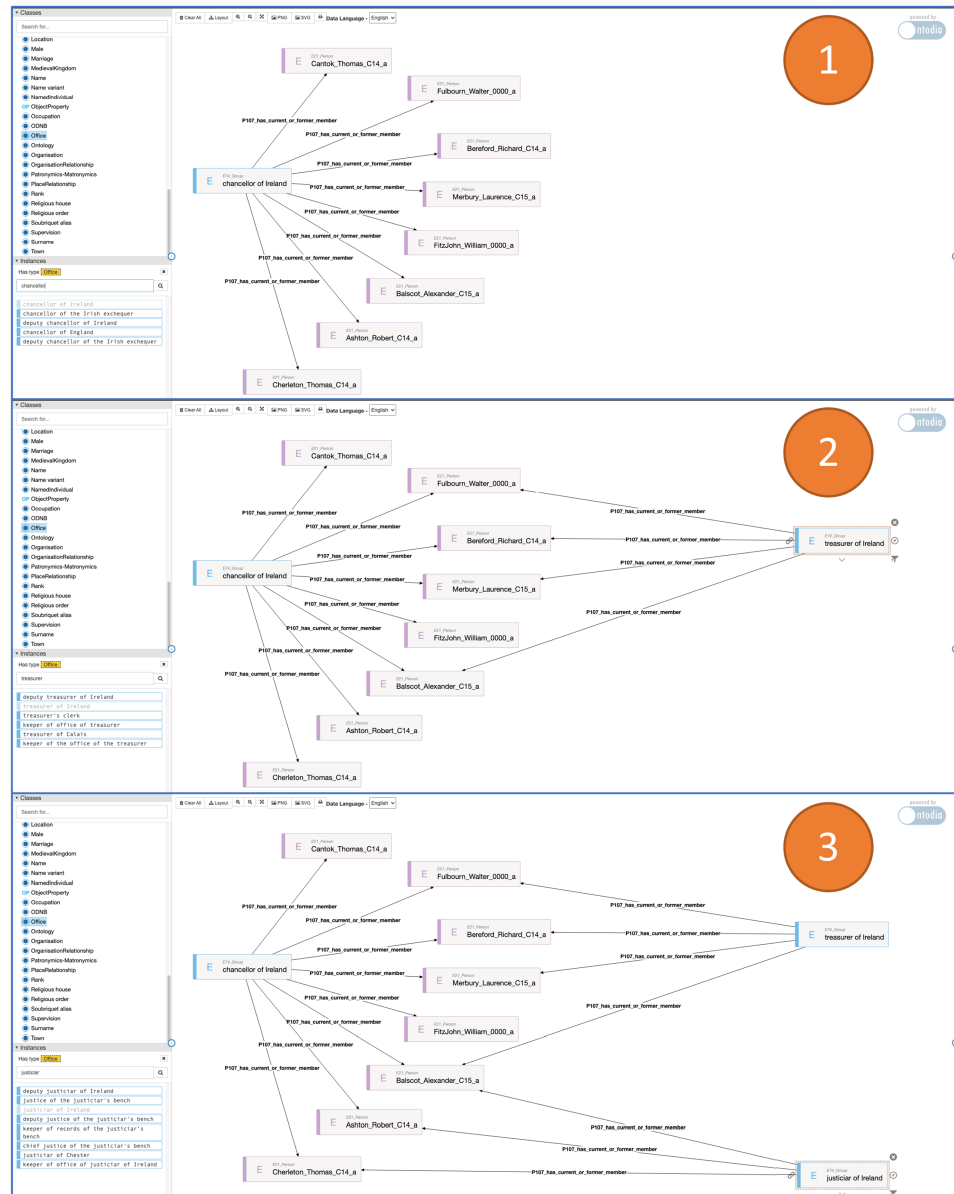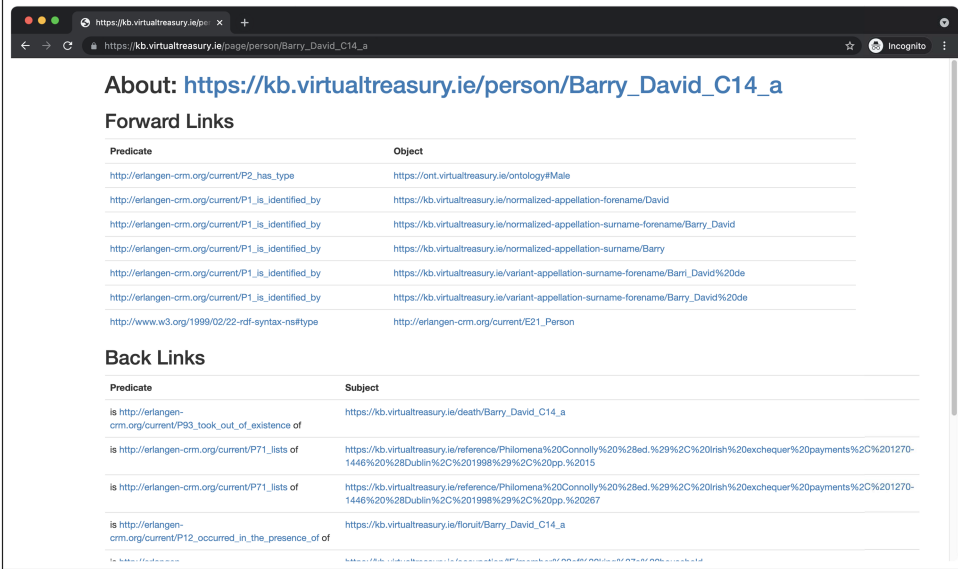
Fig. 11. Visually discovering who was part of the following three offices: "treasurer of Ireland", "chancellor of Ireland", and "justiciar of Ireland" based on the information distilled from the Irish Exchequer Index. We can now easily discover who held multiple positions. The diagrams were rendered with Ontodia.

representations. The Beyond 2022 platform provides a Linked Data frontend[17] allowing historians access to the graphs, provenance graphs, and resources in the triplestores via IRI lookups. The Linked Data frontend provides

---

[17]A bespoke frontend was developed since R2RML prescribes how values should be percent-encoded to ensure so-called IRI-safe IRIs (Section 7.3 of [4]). Existing front ends had difficulty with these percent-encoded values; they tried to decode these values, which resulted in the wrong IRIs being sought in the triplestores.

Fig. 12. Triples displayed on a webpage for a graph. The resource on the right represents a person. Below, there are links allowing one to download the triples in various RDF serialization formats.

a representation based on the content-type passed with the HTTP GET request (e.g., "text/html" for HTML and "text/turtle" for the Turtle RDF serialization format).

Our frontend provides representations for named graphs and resources (see Figures 12 and 13). For the former, we list the triples. For the latter, we provide both forward and backward links.

8 RELATED WORK

In Beyond 2022, subject matter experts ingest factoids they have curated. In our approach, the factoids are kept in different named graphs and these named graphs' provenance information is described using PROV-O. The use of PROV-O and named graphs allows one to assess the different sets of factoids and choose which ones to use. The use of named graphs allows us to cope with subject matter experts' different perspectives of approaches. A slightly different approach was undertaken in the Irish Record Linkage project [5]. In this project, archivists transcribed vital records into one KB. Those records were then interpreted by historians to answer their competency questions by creating a second KB from the first. In their approach, observations by archivists were thus separated from (different) interpretations. While their approach emphasized that separation, their approach more or less corresponds with the use of different graphs.

There are quite a few examples of CIDOC-CRM for Linked Data: museums [3], historical events [25], archeology [36], and biographies [19]. One can observe a number of ways CIDOC-CRM has been extended. Rather than extending CIDOC-CRM, [3] created a new ontology to service the additional information to their users. The work presented in [25] to, like us, to extend CIDOC-CRM. They furthermore recognized the challenges that came with modeling historic places and place names. They avail of a historical place names gazetteer, which we intend to do with Linked Logainm [31] as future work. In our approach, we model the names and geometries of places as properties that have been attributed to places. While such an attribution (an instance of `cidoc:E13_Attribute_Assignment`) does render the KG more complex, it can be easily modeled as an event on which dates can be added, allowing us to easily capture the history and even evolution of a place. In [36], the authors described how they used the model to bridge the "communication barrier" between computer scientists

Fig. 13. Triples displayed on a webpage for a resource. Below, there are links allowing one to download the triples in various RDF serialization formats.

and archeologists. The model does indeed provide a comprehensible upper ontology for the cultural heritage domain. One aspect of CIDOC-CRM that seems counter-intuitive for the Semantic Web is its "type" relationship (cidoc:P2_has_type). That means that arbitrary RDF agents may miss out on some of the types used in this particular KG. This is one of the challenges of CIDOC-CRM identified in [24].

Unlike [32], we have chosen to adopt an upper ontology (CIDOC-CRM) and extend it to fit our needs. In [32], the authors have argued that the upper ontologies existing (in the digital humanities space) were too broad for their purposes. It is true that this exercise may, at times, create many triples to represent simple factoids. We have, however, chosen to adopt an upper ontology, not only to maximize interoperability not only in terms of the resource, but also in terms of the community aware of the CIDOC-CRM ontology, which would have been the downside of developing a bespoke ontology.

There are indeed other initiatives for creating KGs in the digital humanities and cultural heritage space. We first focussed on related work availing of CIDOC-CRM. Others have looked into facilitating data curation in that space. DaCuRa [8], for instance, proposed a niche-sourcing method in which contributors provided facts they encountered in assets. Multiple people contributed facts about the same asset, and one could then curate a final version from all these inputs. Similarly, [37] has developed a software platform (based on a Drupal CMS) on top of a KG that facilitates users in curating, refining, and correcting the RDF statements. In our approach, only a few historians were approached to contribute to the Beyond 2022 KG. As of yet, the involvement of external people has not been considered as we first focus on an authoritative KG to be used within the project. By "authoritative", we mean people approached by the Beyond 2022 project for their subject matter expertise that falls within the project's scope.

9 CONCLUSIONS AND FUTURE WORK

The aim of this article was to present our approach to knowledge organization for Beyond 2022's KG. Novel in our approach is the interdisciplinary approach to historical data management, which we support with KG technologies. We use named graphs for both separating factoids (gathered from different sources, people, and

processes) and representing the mutable nature of historical knowledge. Provenance information about these graphs of factoids allows one to assess its characteristics (e.g., to include or exclude particular graphs from a query). We also proposed two triplestores and endpoints: one for the latest versions of factoids for deployment purposes, and one which keeps track of the provenance and revisions of all graphs of factoids. It is believed that the latter will not only grow much larger (for obvious reasons), but that it will be useful for specific tasks, e.g., assessing the trustworthiness of a set of factoids.

While we have already ingested contemporary places (from the Republic of Ireland) in the KG, future work includes, in the shorter term, the integration of historic places, which will require mapping the evolution of places and their names by relating different attribute assignments. We have also integrated [27], which provides a comprehensive list of all the chief governors of Ireland from the late 12th to early 20th century. The work on [27] has yet to reach the same level of maturity as the Irish Exchequer Index. In the longer term, we will incorporate different sources of information informed by the competency questions and collaboration with the subject matter experts.

REFERENCES

[1] Sebastian Colutto, Philip Kahle, Günter Hackl, and Günter Mühlberger. 2019. Transkribus. A platform for automated text recognition and searching of historical documents. In *Proceedings of the 15th International Conference on eScience.* IEEE, 463–466. DOI:https://doi.org/10.1109/eScience.2019.00060

[2] Philomena Connolly. 1998. *Irish Exchequer Payments 1270-1326.* Irish Manuscripts Commission.

[3] Mariana Damova and Dana Dannells. 2011. Reason-able view of linked data for cultural heritage. In *Proceedings of the 3rd International Conference on Software, Services and Semantic Technologies.* Darina Dicheva, Zdravko Markov, and Eliza Stefanova (Eds.). Springer Berlin, 17–24.

[4] Souripriya Das, Richard Cyganiak, and Seema Sundara. 2012. *R2RML: RDB to RDF Mapping Language.* W3C Recommendation. W3C. Retrieved October 28, 2020 from https://www.w3.org/TR/2012/REC-r2rml-20120927/.

[5] Christophe Debruyne, Oya Deniz Beyan, Rebecca Grant, Sandra Collins, Stefan Decker, and Natalie Harrower. 2016. A semantic architecture for preserving and interpreting the information contained in Irish historical vital records. *International Journal on Digital Libraries* 17, 3 (2016), 159–174. DOI:https://doi.org/10.1007/s00799-016-0180-8

[6] Christophe Debruyne, Gary Munnelly, Lynn Kilgallon, Declan O'Sullivan, and Peter Crooks. 2020. *Beyond 2022 Knowledge Graph Sample Data.* Retrieved November 15, 2020 from https://doi.org/10.5281/zenodo.4276353

[7] Christophe Debruyne and Declan O'Sullivan. 2016. R2RML-F: Towards sharing and executing domain logic in R2RML mappings. In *Proceedings of the Workshop on Linked Data on the Web, LDOW 2016, Co-located with 25th International World Wide Web Conference* .Sören Auer, Tim Berners-Lee, Christian Bizer, and Tom Heath (Eds.). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1593/article-13.pdf.

[8] Kevin Chekov Feeney, Declan O'Sullivan, Wei Tai, and Rob Brennan. 2014. Improving curated web-data quality with structured harvesting and assessment. *International Journal on Semantic Web Information Systems* 10, 2 (2014), 35–62. DOI:https://doi.org/10.4018/ijswis.2014040103

[9] Michael Fewer. 2019. The battle of the four fourts, 28–30 June 1922. *History Ireland* 27, 4 (2019), 44–47. Retrieved from https://www.jstor.org/stable/26853089.

[10] Daniel Garijo. 2017. WIDOCO: A wizard for documenting ontologies. In *Proceedings of the 16th International Semantic Web Conference.* Claudia d'Amato, Miriam Fernández, Valentina A. M. Tamma, Freddy Lécué, Philippe Cudré-Mauroux, Juan F. Sequeda, Christoph Lange, and Jeff Heflin (Eds.), Lecture Notes in Computer Science, Vol. 10588, Springer, 94–102. DOI:https://doi.org/10.1007/978-3-319-68204-4_9

[11] Günther Goerz, Martin Oischinger, and Bernhard Schiemann. 2008. An implementation of the CIDOC conceptual reference model (4.2.4) in OWL-DL. In *Proceedings of the 2008 Annual Conference of CIDOC-The Digital Curation of Cultural Heritage.*

[12] José Manuél Gómez-Pérez, Jeff Z. Pan, Guido Vetere, and Honghan Wu. 2017. Enterprise knowledge graph: An introduction. In *Proceedings of the Exploiting Linked Data and Knowledge Graphs in Large Organisations.* Jeff Z. Pan, Guido Vetere, José Manuél Gómez-Pérez, and Honghan Wu (Eds.). Springer, 1–14. DOI:https://doi.org/10.1007/978-3-319-45654-6_1

[13] Tom Gruber. 2009. *Ontology.* Springer US, Boston, MA, 1963–1965. DOI:https://doi.org/10.1007/978-0-387-39940-9_1318

[14] Michael Grüninger and Mark S. Fox. 1995. *The Role of Competency Questions in Enterprise Engineering.* Springer US, Boston, MA, 22–31. DOI:https://doi.org/10.1007/978-0-387-34847-6_3

[15] Steven Harris and Andy Seaborne. 2013. *SPARQL 1.1 Query Language.* W3C Recommendation. W3C. Retrieved October 28, 2020 from https://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

[16] Olaf Hartig and Jun Zhao. 2010. Publishing and consuming provenance metadata on the web of linked data. In *Proceedings of the Provenance and Annotation of Data and Processes—3rd International Provenance and Annotation Workshop.* Deborah L. McGuinness, James

Michaelis, and Luc Moreau (Eds.), Lecture Notes in Computer Science, Vol. 6378, Springer, 78–90. DOI:https://doi.org/10.1007/978-3-642-17819-1_10

[17] Martin Hepp. 2008. Ontologies: State of the art, business potential, and grand challenges. In *Proceedings of the Ontology Management, Semantic Web, Semantic Web Services, and Business Applications*. Martin Hepp, Pieter De Leenheer, Aldo de Moor, and York Sure (Eds.). Semantic Web and Beyond: Computing for Human Experience, Vol. 7. Springer, 3–22. DOI:https://doi.org/10.1007/978-0-387-69900-4_1

[18] Rinke Hoekstra and Paul Groth. 2014. PROV-O-Viz—understanding the role of activities in provenance. In *Proceedings of the Provenance and Annotation of Data and Processes—5th International Provenance and Annotation Workshop*. Bertram Ludäscher and Beth Plale (Eds.), Lecture Notes in Computer Science, Vol. 8628. Springer, 215–220. DOI:https://doi.org/10.1007/978-3-319-16462-5_18

[19] Eero Hyvönen, Miika Alonen, Esko Ikkala, and Eetu Mäkelä. 2014. Life stories as event-based linked data: Case semantic national biography. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference*. Matthew Horridge, Marco Rospocher, and Jacco van Ossenbruggen (Eds.), Vol. 1272. CEUR-WS.org, 1–4. Retrieved from http://ceur-ws.org/Vol-1272/paper_5.pdf.

[20] Mustafa Jarrar and Robert Meersman. 2009. Ontology engineering—the DOGMA approach. In *Proceedings of the Advances in Web Semantics I - Ontologies, Web Services and Applied Semantic Web*. Tharam S. Dillon, Elizabeth Chang, Robert Meersman, and Katia P. Sycara (Eds.), Lecture Notes in Computer Science, Vol. 4891. Springer, 7–34. DOI:https://doi.org/10.1007/978-3-540-89784-2_2

[21] Holger Knublauch and Dimitris Kontokostas. 2017. *Shapes Constraint Language (SHACL)*. W3C Recommendation. W3C. Retrieved October 28, 2020 from https://www.w3.org/TR/2017/REC-shacl-20170720/.

[22] Mikko Koho, Erkki Heino, and Eero Hyvönen. 2016. SPARQL faceter - client-side faceted search based on SPARQL. In *Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop Co-located with the 13th Extended Semantic Web Conference ESWC 2016*. Raphaël Troncy, Ruben Verborgh, Lyndon J. B. Nixon, Thomas Kurz, Kai Schlegel, and Miel Vander Sande (Eds.), Vol. 1615. CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1615/semdevPaper5.pdf.

[23] Timothy Lebo, Deborah McGuinness, and Satya Sahoo. 2013. *PROV-O: The PROV Ontology*. W3C Recommendation. W3C. Retrieved July 14, 2020 from https://www.w3.org/TR/2013/REC-prov-o-20130430/.

[24] Chia-Hung Lin, Jen-Shin Hong, and Martin Doerr. 2008. Issues in an inference platform for generating deductive knowledge: A case study in cultural heritage digital libraries using the CIDOC CRM. *International Journal on Digital Libraries* 8, 2 (2008), 115–132. DOI:https://doi.org/10.1007/s00799-008-0034-0

[25] Eetu Mäkelä, Juha Törnroos, Thea Lindquist, and Eero Hyvönen. 2017. WW1LOD: An application of CIDOC-CRM to world war 1 linked data. *International Journal on Digital Libraries* 18, 4 (2017), 333–343. DOI:https://doi.org/10.1007/s00799-016-0186-2

[26] Lucy McKenna, Christophe Debruyne, and Declan O'Sullivan. 2018. Understanding the position of information professionals with regards to linked data: A survey of libraries, archives and museums. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. Jiangping Chen, Marcos André Gonçalves, Jeff M. Allen, Edward A. Fox, Min-Yen Kan, and Vivien Petras (Eds.). ACM, 7–16. DOI:https://doi.org/10.1145/3197026.3197041

[27] T. W. Moody, F. X. Martin, and F. J. Byrne. 2011. *A New History of Ireland, Volume IX: Maps, Genealogies, Lists: A Companion to Irish History, Part II Illustrated Edition A New History of Ireland, Volume IX: Maps, Genealogies, Lists: A Companion to Irish History, Part II Illustrated Edition*. Oxford University Press.

[28] Dmitry Mouromtsev, Dmitry Pavlov, Yury Emelyanov, Alexey Morozov, Daniil Razdyakonov, and Mikhail Galkin. 2015. The simple web-based tool for visualization and sharing of semantic data and ontologies. In *Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference*. Serena Villata, Jeff Z. Pan, and Mauro Dragoni (Eds.). Vol. 1486. CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1486/paper_77.pdf.

[29] Yves Raimond and Guus Schreiber. 2014. *RDF 1.1 Primer*. W3C Note. W3C. Retrieved July 14, 2020 from https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/.

[30] Zoë Reid. 2018. Unwrapping the Past: Conserving Archives Damaged in the Fire That Destroyed the Public Record Office of Ireland. Retrieved July 15, 2021 from https://beyond2022.ie/wp-content/uploads/2019/01/9-Unwrapping-the-past.-Zoe-Reid.pdf.

[31] Catherine Ryan, Rebecca Grant, Eoghan Ó Carragáin, Sandra Collins, Stefan Decker, and Nuno Lopes. 2015. Linked data authority records for Irish place names. *International Journal on Digital Libraries* 15, 2-4 (2015), 73–85. DOI:https://doi.org/10.1007/s00799-014-0129-8

[32] Cogan Shimizu, Pascal Hitzler, Quinn Hirt, Dean Rehberger, Seila Gonzalez Estrecha, Catherine Foley, Alicia M. Sheill, Walter Hawthorne, Jeff Mixter, Ethan Watrall, Ryan Carty, and Duncan Tarr. 2020. The enslaved ontology: Peoples of the historic slave trade. *Journal of Web Semantics* 63 (2020), 100567. DOI:https://doi.org/10.1016/j.websem.2020.100567

[33] Peter Spyns, Robert Meersman, and Mustafa Jarrar. 2002. Data modelling versus ontology engineering. *SIGMOD Record* 31, 4 (2002), 12–17. DOI:https://doi.org/10.1145/637411.637413

[34] Ed Summers and Antoine Isaac. 2009. *SKOS Simple Knowledge Organization System Primer*. W3C Note. W3C. Retrieved July 14, 2020 from https://www.w3.org/TR/2009/NOTE-skos-primer-20090818/.

[35] Jeni Tennison. 2016. *CSV on the Web: A Primer*. W3C Note. W3C. Retrieved July 15, 2021 from https://www.w3.org/TR/2016/NOTE-tabular-data-primer-20160225/.

[36] Stefano Valtolina, Piero Mussio, Giovanna Gianni Bagnasco, Pietro Mazzoleni, Stefano Franzoni, Muriel Geroli, and Cristina Ridi. 2007. Media for knowledge creation and dissemination: Semantic model and narrations for a new accessibility to cultural heritage. In *Proceedings of the 6th Conference on Creativity & Cognition.* Ben Shneiderman, Gerhard Fischer, Elisa Giaccardi, and Michael Eisenberg (Eds.). ACM, 107–116. DOI:https://doi.org/10.1145/1254960.1254976

[37] Gemma Webster, Hai H. Nguyen, David E. Beel, Chris Mellish, Claire D. Wallace, and Jeff Z. Pan. 2015. CURIOS: Connecting community heritage through linked data. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing.* Dan Cosley, Andrea Forte, Luigina Ciolfi, and David McDonald (Eds.). ACM, 639–648. DOI:https://doi.org/10.1145/2675133.2675247

[38] Jun Zhao and Olaf Hartig. 2012. Towards interoperable provenance publication on the linked data web. In *Proceedings of the WWW2012 Workshop on Linked Data on the Web.*