WP9-D92

Solution Planning Technology

Project full title:	Knowledge Driven Data Exploitation
Project acronym:	K-Drive
Grant agreement no.:	286348
Project instrument:	EU FP7/Maria-Curie IAPP/PEOPLE WP 2011
Document type:	D (deliverable)
Nature of document:	R (report)
Dissemination level:	PU (public)
Document number:	UNIABDN, ESI, ISOCO/WP9-D92/D/PU/b1
Responsible editors:	Jeff Z. Pan, Ronald Denaux, Hai Nguyen, Jose Manuel Gomez Perez, Panos Alexopoulos
Reviewers:	Yuting Zhao
Contributing participants:	UNIABDN, ESI, ISOCO
Contributing workpackages:	WP9
Contractual date of deliverable:	30 June 2016
Actual submission date:	30 June 2016

Abstract

Keyword List
Guidance, Interactive UI, Vagueness Annotation, Linked Open Vocabulary Discovery, Personalized Dataset Summarization

Project funded by the European Commission within the 7th Framework Programme/Maria Curie Industry-Academia Partnerships and Pathways schema/PEOPLE Work Programme 2011.

Solution Planning Technology

Jeff Z. Pan[1], **Ronald Denaux**[2], **Hai Nguyen**[1], **Jose Manuel Gomez Perez**[2], **Panos Alexopoulos**[3]

[1] Department of Computing Science, Aberdeen University, UK
[2] Expert System Iberia, Spain
[3] iSOCO, Spain

30 June 2016

Abstract

Keyword List
Guidance, Interactive UI, Vagueness Annotation, Linked Open Vocabulary Discovery, Personalized Dataset Summarization

Contents

1 Introduction

In this work package (WP9) we further extend the intelligent User Interface (WP4 & WP7), not only letting people browse and knowledge base, constructing queries (WP3 & WP5) and exploring the query results (WP6), but also allow people to build complex tasks and generate plans to help people undertake the tasks to solve problems.

Knowledge Graph (KG) is selevted as the core technique to understand the user's intention and the complex tasks required by users, as well as to gain the accurate summarizations of the datasets.

It will be interesting to see how knowledge graphs have been used in real-life scenarios, what benefits these techniques can bring and what we can learn from these applications. To answer these questions, in this deliverable, we present success stores of the applications of knowledge graph techniques from various domains (healthcare, media and culture) and different organisations (international company - IBM, Small and Medium Enterprise - HAVAS, and University - the University of Aberdeen).

2 Applying Knowledge Graphs in Healthcare

Here we present an application of leveraging knowledge graphs in healthcare application. In this application we deal with discovering clinical appropriateness in oncology, based on technologies and services of knowledge graphs.

This complicated topic has a big impact on what is called *science/evidence-based medicine*, therefore, on the role of real data and on real knowledge implemented in clinical decision support systems. The final aim is to design technologies and solutions that may empower clinicians in their complex decision-making processes. In particular, we aim to present the role of knowledge graphs implemented within a clinical decision support system.

The practice of medicine requires the integration of vast and continuously changing information for the prescription of appropriate treatments, therefore, the availability of a powerful tools like dynamic knowledge graphs represent a significant advantage when designing decision support systems. Bridging together several information sources is not that simple as it may look like; in particular, when those sources may contain not fully consistent information and whose quality can be compromised because of data entry mistakes, having a knowledge graph approach can be very beneficial to address so many practical issues.

2.1 The Problem in Clinical Practice Guidelines

It is well known that diseases are not only major sources of human suffering and one of the most common causes of death worldwide, but also bring heavy financial impact to human society. For example cancer is one of the most common diseases, besides its tremendous health impact, cancer also bears a staggering financial burden on the world's economy reached $ 895 billion, accounting for approximately 1.5% of the world's GDP in 2008.

In fact the immense efforts invested to develop cancer treatments have gradually lowered cancer mortality rates, but have also complicated the process of oncology care. There are now a plethora of cancer treatment options, making it a challenging task to consistently follow the treatment recommendations dictated by *Clinical Practice Guidelines* (CPGs). Interestingly, although deviations from guidelines can have negative results, they are often in fact beneficial. Thus, a central challenge in cancer medical informatics is to identify deviations from CPGs [1] and to assess whether they are medical mistakes or guideline improvements.

The modern medical landscape is characterised by a plethora of different treatment options for even almost indistinguishable clinical statuses. While the development of new treatment modalities is beneficial, it also poses challenges associated with the growing body of evidence regarding the outcomes of different treatments.

As a consequence of the complexity of treatment possibilities and the presence of widespread variation in medical practice, it has become clear that a large fraction of patients do not in fact receive the best possible care [1,2]. Deviations from optimal care are abundant in diseases where treatment efficacy varies as a result of subtle changes in the clinical scenario as well as in cases where clear scientific evidence is not present, as is often seen in cancer [3,4]. Therefore, an important question in medicine is what leads clinicians to prescribe treatments that do not adhere to best practice.

One approach to monitor deviations from standard medical practice is by assessing adherence to CPGs.

CPGs are promoted as a means to decrease inappropriate practice variation and reduce medical errors [6]. It is generally thought that clinician adherence to CPG recommendations is the primary means to achieve this goal. High levels of adherence to CPGs may indicate optimal care, whereas low adherence rates may suggest sub-optimal treatment. In reality, however, deviation from CPGs often reflects the fact that CPGs cannot be exhaustive; it is not feasible to cover the entire combinatorial space of patient parameters. Deviations from CPG recommendations may thus be beneficial, and it is expected that clinicians will use their personal judgement to contextualise individual patient decisions. In light of the above, previous work identified several barriers to adherence including physician familiarity with the CPGs, physician attitudes towards the CPGs, environmental factors, CPG implementation factors and patient-related factors such as preference [7,8].

Monitoring compliance to CPGs in the clinical setting can be labor intensive. Therefore, in this study we strived to automate the characterisation of adherence to CPGs using natural language processing, data modelling and comparison algorithms. Our vision was to computationally parse *Electronic Health Records* (EHRs) containing both structured and unstructured data to quantify adherence levels, categorise the types of deviations from CPG recommendations, and finally identify the potential rationale for these deviations. We demonstrate our approach using the EHRs of patients diagnosed with adult *Soft-Tissue Sarcoma* (STS). STS is a group of connective-tissue based cancers that account for roughly 1survival rates of slightly greater than 50have diverse anatomical origins and can derive from multiple somatic cell types. The variety of histologies results in the presence of multiple drug options and the different anatomical locations offer multiple surgical possibilities. As STSs are rare cancers with numerous treatment options, it is not surprising that prescriptions for patients frequently deviate from CPGs, making STS an ideal use case to evaluate our methodology [10,11,12].

2.2 Preparing the Data and Building the Knowledge Graphs

Data stored within the hospital medical records systems are complex, heterogeneous and stratified over time, and not even necessarily organised by following an inter-temporal coherent and effective policy. Clinicians when they have to make decisions they need to face a lot of complexity to explore and integrate a wide range of information. Machines, or decision support systems, they also need a plethora of data that need to be qualified and integrated in a complete, accurate and consistent knowledge base

[1] CPGs are collective sets of treatment recommendations that attempt to capture the best medical practices for different pathologies [5].

so that useful insight can be extracted out of that. Data curation is therefore extremely important before designing any further services based on top of them and again, the knowledge graph approach is extremely valuable.

Looking at the data, originally they are about clinical tests, patients' medical history and clinical status, guidelines, clinical literature, etc. Specially efforts were made in order to prepare the proper data for building knowledge graphs, which including applying *Natural Language Processing* (NLP) technologies.

In the knowledge graphs, all the data structure is patient-centred and integrates information dealing with standard information such as demographics, clinical status, oncological disease description but also non oncological information about the patients' eventual comorbidities, etc.

2.2.1 Description of concepts

The CPGs used in this study were developed by the *Lombardy Oncology Network*, a data sharing network that contains over fifty care premises in Northern Italy. Patient data used in this work was gathered at the *Fondazione IRCCS Istituto Nazionale dei tumouri* (INT), a network member and thought leader, from November 2006 to November 2012.

The CPGs contained hundreds of clinical cancer presentations (conceptually similar to diagnosis) with their matching recommended treatments. There are multiple recommended treatments for each clinical presentation. A single CPG recommendation was defined as unique coupling of clinical presentation, recommended treatment, and start/end date. The study involved 1484 separate CPG recommendations.

Individual clinical presentations were modelled as a data structure of the following clinical fields: tumour anatomic location, tumour depth (deep/superficial), tumour grade, tumour size, disease status, tumour histological type (liposarcoma, etc.) and surgical status (tumour resectable/not resectable). A clinical presentation can include all or a subset of the fields. This modelling approach is standard for CPGs, and is similar to that used by the National Comprehensive Cancer Network [13]. An example of an STS clinical presentation in the Lombardy CPGs is: “Patient with adult soft tissue sarcoma located in the limb or torso with a deep, high grade, =5cm, localised tumour”.

Treatment programs (TPs) were defined as sequences of medical procedures (treatment elements), for example “Wide surgical excision with adjuvant/neo-adjuvant radiotherapy”. A treatment element can contain items such as drug administration, surgery, radiotherapy, and transplantation. The Lombardy CPGs contained recommended TPs for each clinical presentation.

Once a physician selected a particular clinical presentation from the CPGs, the matching TPs were presented via the local EHR system. Physicians were entitled to prescribe a TP that was discordant with CPG recommendations (Figure1). In doing so, they subsequently detailed the contents of their alternative prescribed TP in free-text form. The EHR system recorded this decision as well as additional relevant notes provided by the physicians.

Data regarding the treatment was also entered into the EHR system by caregivers during TP execution. We applied standard NLP methods on this data to deduce the actual TP that a patient underwent. The extracted TP was compared to the CPG recommended TPs to assess adherence. The actual TP was considered to deviate if it was discordant to CPG recommendations, regardless of whether the prescription was according to the recommended TPs or not (Figure1).

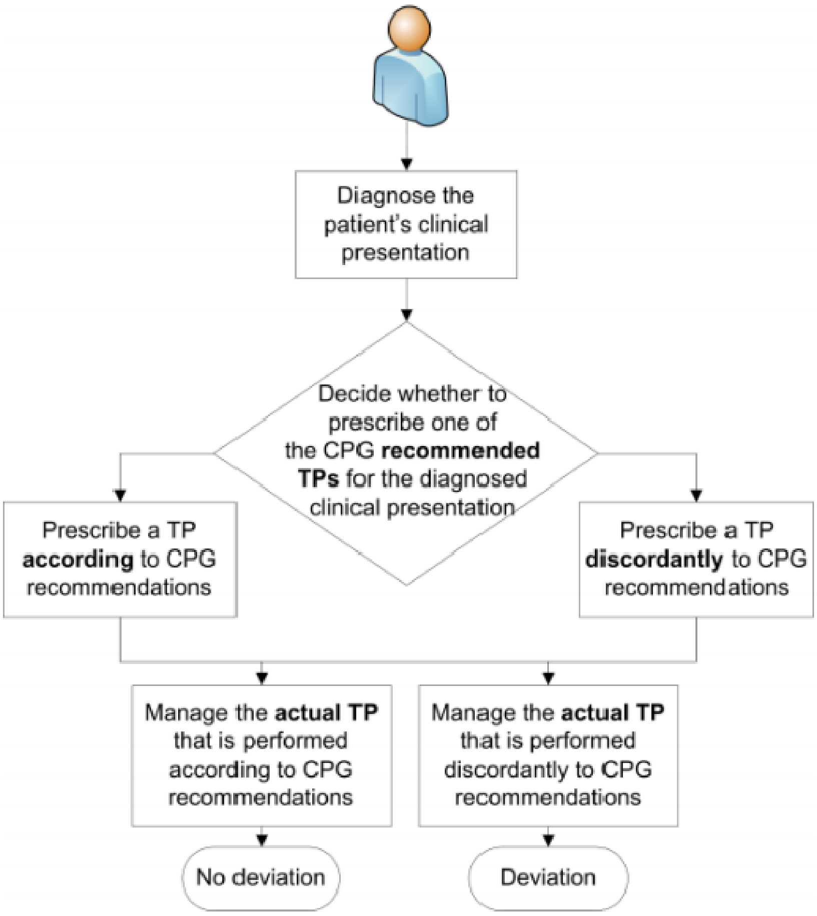

Diagnose the patient's clinical presentation
Decide whether to prescribe one of the CPG recommended TPs for the diagnosed clinical presentation
Prescribe a TP according to CPG recommendations
Prescribe a TP discordantly to CPG recommendations
Manage the actual TP that is performed according to CPG recommendations
Manage the actual TP that is performed discordantly to CPG recommendations
No deviation
Deviation

Figure 1: CPG assisted decision making

2.2.2 Application of NLP Technique on Electronic Health Records

We applied NLP techniques on the EHR free text data to computationally retrieve the required information for this study. After Italian to English machine translation, we used the *Unstructured Information Management Architecture* (UIMA) framework to process unstructured information [14]. Our UIMA pipeline included tokenisation, *parts of speech* (POS) tagging, normalisation using standard terminologies in UMLS [15], entity and relationship extraction, semantic analysis, negation and disambiguation reasoning. Resulting structured annotations included drugs, diseases, procedures, symptoms, body regions and tumour characteristics. Relationship extractions were used to infer aspects such as the number of chemotherapy cycles, tumour size, tumour grade, and reasons for specific treatment prescription.

2.2.3 Study setting, patient selection, and data cleansing

Our patient data encompassed adult STS patients treated at the Fondazione IRCCS Istituto Nazionale dei tumori between November 2006 and November 2012. We acquired 5598 electronic patient discharge letters representing 2699 STS treatment programs on a total of 2151 different patients. 948 TPs with missing data were excluded consisting of TPs that were follow-ups, the actual TP was unknown, did not have at least one CPG recommendation due to CPG incompleteness, or were clinical studies not mentioned in the CPGs. This resulted in 1751 TPs consisting of 1431 patients on whom we performed the analysis presented in this study. Some patients had two or more sequentially prescribed TPs.

2.3 Services based on the Knowledge Graphs

Because of the dynamic of the knowledge graphs in this use case, the services should be time-featured, in the sense that the services are dynamic. Based on the decision models, the services can be classified into: discovering deviations, classification of deviations, and justification (explanation) about above results.

2.3.1 Treatment program comparison

The results of the text analytics are structured annotations on the text. The annotations first need to be transformed to a pre- defined data model to enable advanced analyses. We therefore designed an actual TP model that defines the treatment which was given to patients. The model was designed to enable comparison with the recommended TPs.

To categorise deviations we identified the most similar recommended TP in the CPGs. The most similar recommended TP was found by assessing the degree of similarity between recommended and actual TPs. The differences between a deviating actual TP and its most similar recommended TP were classified into different categories.

2.3.2 Extracting reasons for deviation

We used the same NLP techniques described above to extract reasons for deviation from CPGs. This was done by identifying relationships between extracted annotations using semantic parsing rules. For example, one can consider the following machine-translated sentence: "In light of extension of illness, the patient's age and preliminary activity of molecule in this particular histotype, starting chemotherapy with gemcitabine". By detecting that the conjunction "in light of" connects the two parts of the sentence, we deduced that the first part of the sentence describes reasons for the given treatment, whereas the second part ("starting chemotherapy ") describes the treatment itself.

2.3.3 Manual validation

We performed manual validation of our computational results on a subset of randomly selected TPs. Four human validators were exposed to the entire EHR records and CPGs. Different subsets of the validation dataset were allocated to each reviewer and results were compiled.

2.3.4 Quantifying factors that impact deviation frequency

Our computational approach identified deviations in 48.9% of the actual TPs, meaning that the actual given treatments were found to not fully comply with the CPG recommended TPs.

We next assessed non-clincial parameter correlation with deviation frequency. Strikingly, 35% of the TPs prescribed according to CPG recommendations in reality deviated from the CPG recommendations. More expectedly, TPs that were prescribed discordantly to CPG recommendations did in fact deviate in 80% cases. Gender and age (cutoff set at median age) were not associated with deviation frequency.

Upon analysis of clinical parameters, we observed that all disease and tumour parameters were associated with deviation frequencies, except for tumour location. This analysis portrays an expected trend in which poorer prognostic status (large, high grade, and deep tumours) is linked to substantially higher deviation levels. Indeed, the highest deviation frequency was found in metastatic disease (78%).

2.3.5 Measuring prevalence of different deviation types

TP deviations can be classified using non-mutually exclusive categories. It shows the abundance of deviations that have added or removed treatment elements, differences in chemotherapy drugs, differences in number of chemotherapy cycles, and differences in surgery type.

The most abundant source of deviation was over-treatment, consisting of 39.7% of all cases in contrast to 12.7% for missing treatments. Notably, metastatic presentations had no excluded elements despite an overall average of 12.7% for all TPs. Also prominent was the observation that there was only a 10.3% deviation rate of type 'different chemotherapy drug' for metastatic cases with administered chemotherapy. In general, disease parameters were more strongly associated with chemotherapy differences than surgical differences, with an exception being local/metastatic clinical presentations.

2.3.6 Identification of potential reasons for deviation

NLP parsing identified 1191 potential reasons for deviation among the 857 TPs that we labeled as deviations (average 1.4 per TP). 67.3% of the deviating TPs had one to four reasons and 29.7% had no identified reasons.

Potential reasons for deviation were classified into five categories: cancer status, other clinical, current treatment related, previous treatment related, and patient preference related. Reasons for deviation that were based on cancer status represented the majority (59%) of all deviations.

Deviation reasons were further classified into lower-level categories. The cancer status category consisted of different tumour and disease progression parameters. Other clinical related reasons included demographics, oncological and non-oncological comorbidities, acute symptoms and overall clinical condition. The previous treatment related reasons include much previous treatment, poor previous response, previously severe side effects, and presence of residual margins after surgery. The patient preferences category included patient treatment requests or refusals. Lastly, current treatment related reasons consisted of anticipated treatment efficacy, impact on quality of life, and newly available clinical evidence. Deviations due to environmental constraints including lack of personnel or resources were rare and thus not presented.

The largest fraction of deviations appear to result from disease progression or a lack thereof together with presence of acute symptoms. Interestingly, new medical knowledge was only a small fraction of potential deviation causes.

2.4 Contributions to Healthcare Practise

In this section, we introduce a real world application for deep understanding of adherence to CPG recommendations by applying knowledge graphs, used in an adult STS clinical studies. The resulting insights from this work significantly contribute to Healthcare Practise on the following aspects:

- **Helping the doctors to understand clinical deviations**. It identifies deviations, classify them by types, and finally proposes reasons that may reflect the physicians rationale in deviation cases.
- **Helping to understand the decision making process of physicians**. It provides explanations to the technical directors and managers, on why the deviations is so high. Some reasons could be (i) the comorbidities is high, (ii) malpractice, and (iii) CPG is outdated, etc.
- **Improving CPGs**. It can identify cases where deviations may be beneficial, and to increase adherence to CPGs when deemed appropriate. In some situation it can also suggest clinical trials in order to improve the CPGs.

Beyond value in understanding clinical deviations, this current analysis raises multiple observations that may be useful to sarcoma researchers and the decision support community.

3 A Knowledge Graph for Innovation in the Media Industry

3.1 The Business Problem

The communication between brands and consumers is set to explode. Product features are no longer the key to sales and the combination of both personal and collective benefits is becoming an increasingly crucial aspect. As a matter of fact, brands providing such value achieve higher impact and consequently derive clearer economic benefits. On the other hand, millennials are taking over; inducing a dramatic change in the way consumers and brands engage and what channels and technologies are required to enable the process. As a result traditional boundaries within the media industry are being stretched and new ideas, inventions, and technologies are needed to keep up with the challenges raised by the increasing demands of this data-intensive, in-time, personalised, and thriving market.

Thus, it is necessary to leverage advances in the area by stimulating a collaboration ecosystem between the different players. Inspiring examples include the adoption by Tesla Motors of the open patents policy, whereby Tesla shares their innovation in regards to electric cars openly via the internet. In return, Tesla expects the industry to further evolve the electric car and dynamise the market. In the media industry a paradigmatic case of this *better together* approach is HAVAS 18 Innovation Labs, deployed at strategic locations around the world. One of such locations is the Siliwood research centre in Santa Monica, co-created with Orange, which focuses on the convergence between technology, data science, content and media. 18 Innovation Labs seeks to connect a great mix of local talent over the sites, involving innovators, universities, start-ups and technology trends to co-create initiatives relevant now and in the mid-term for both HAVAS and their clients to stay one step ahead.

To achieve that, HAVAS has created an enterprise knowledge graph and information platform that aggregates all the available knowledge about technology startups worldwide and makes it available for exploitation by media business strategists through a single entry point. To the best of our knowledge this is one of the first applications of knowledge graph principles in the enterprise world, and the first in the media industry, after internet search giants Google, Yahoo, and Bing coined the term at web-scale, each with their own implementations. Related initiatives include domain-specific efforts like, social graphs like Facebook, and reference resources like DBpedia and Freebase.

3.2 The HAVAS 18 Knowledge Graph

In a way analogous to the above-mentioned initiatives, the main objective of HAVAS 18 knowledge graph is to enable knowledge-based services for search, discovery and understanding of information about relevant startups in their first 18 months. So, we aimed at providing a unified knowledge graph where:

1. Entities are uniquely identified by URIs and interlinked across sources.
2. Such entities are relevant to HAVAS 18 Labs, including startups, people of interest related to them through different roles, e.g. founder, investor, etc., bigger and more established companies, universities, and technology trends.
3. Rich information is provided about entities (facts, relationships and features).

To this purpose, the graph follows the lifecycle described in chapter 4, comprising three main phases: knowledge acquisition, integration, and consumption. We extract data from online sources, including generalist and specialised web sites, online news, entrepreneurial and general purpose social networks, and other content providers. By maximising the use of sources offering web APIs, we expect to minimise additional unstructured data processing time and complexity, at the cost of unexpected changes in the APIs, a potential source of decay in the knowledge graph. Data sources include:

- **Core data** from specialised sites like AngelList and CrunchBase, with useful facts about the main entities in the graph (startups, innovators, investors, other companies and universities), the relationships between them, and domain-specific news.
- **Relationships:** Beyond factual knowledge about the entities, the resulting graph makes emphasis on how they are related to each other. We enrich the relationship graph with information from Facebook, LinkedIn, and Twitter which helps completing a social and professional graph between the entities. Such explicit relationships support the discovery of new insights and navigation.
- **Extended media coverage,** with general news coming from media in any domain through Newsfeed.ijs.si. News text is processed with the semantic annotation framework Knowledge Tagger (see section 4.4) in order to resolve entities and disambiguate.

The data are structured and integrated in an RDF dataset. The underlying schema is built on top of a number of standard W3C vocabularies, including Schema.org, FOAF, and SKOS, and IPTC's rNews. On top of the data, a service layer is provided through a RESTful API with JSON and API key authentication. The API allows the exploitation of the graph by application developers and ultimately media business strategists through analytics platforms and dedicated user interfaces. API services include CRUD methods for entity and relationship management, graph navigation, search, and definition and access to business KPIs about the startups.

In addition to automated information extraction means, the knowledge graph can also be populated with on-site information by local rapporteurs, members of the local entrepreneurial scene distributed at each of the HAVAS 18 Innovation Labs. Rapporteurs are provided with means to add or modify entities and relationships in the graph, following the schema, assisted by autocomplete functionalities that leverage the knowledge previously stored in the graph. They also play the role of curators of knowledge produced either by other peer rapporteurs or extracted automatically. The combination of automatic methods and human expertise allow that a common knowledge graph can be leveraged consistently across the company.

Currently, the graph contains information about 1.812 startups, 559 technology trends, 1.597 innovators, 20 companies and 35 universities and research centres in Siliwood, following the Linked Data principles. All these entities are additionally connected to relevant online news, where they are mentioned (currently, 36.802), for extended and up-to-date information about them. The Knowledge Graph is updated daily in an automated batch process, identifying new entities and updating existing ones. We expect the knowledge graph to quickly reach the threshold of 300.000 startups below 18 months and extend to the remaining Labs in the next few months.

3.3 Value Proposition

Innovation is often misunderstood and difficult to integrate into companies' mindset and culture. So, why not activate relevant external talent and resources when necessary? The discovery and surveillance of trends and talent in the start-up ecosystem can be time consuming, though. HAVAS's knowledge graph sets its semantic engineering to run a surveillance monitoring of the entrepreneurial digital footprint, collecting and gathering fruitful insight and information, which provides the Innovation Labs staff with clear leads for their analyses. This is the key approach and philosophy of HAVAS. By automating part of the research process, it can get there faster and more accurately than competitors, leveraging millions of data points, and implementing consistency through a single and shared knowledge entry point.

At the moment this assisted process is integrated with a manual audit of trends and start-ups, executing a series of evaluation matrices to weigh and assess each individual entity in the graph against HAVAS' business needs. The knowledge graph is being opened to HAVAS' network, with teams in 120 offices around the world and clients, providing access to knowledge about best-in-class talent to implement new thinking and cutting edge solutions to the never ending and evolving challenges within the media industry. Based on the knowledge graph, teams also rate and share experiences, ensuring that learning can be propagated across the network.

3.4 Challenges

To optimise the trustworthiness and accuracy of the graph, we maximised the use of authoritative and specialised sources and prioritised freshness over volume. However, entity resolution and disambiguation is an issue, especially when unstructured data from unbounded domains come into play. During data integration and enrichment, several candidate entities can be identified. In order to resolve the correct one we defined evidence models on top of the schema with the key classes required to provide an entity class with univocal context information. For example, in the case of Startup, this could be Founder, Client, and Technology. Upon data harvesting, e.g. from news, the text is processed by Knowledge Tagger, which extracts entities and matches them against the evidence model providing a measure of evidence based on the context fragments identified in the text and allowing ranking. Further complexity is added when an entity which is not in the domain of interest, e.g. Domo, the gas station, has to be discriminated from one which is part of such domain and potentially in the graph, e.g. Domo, the startup.

Other challenges include (sub)graph time and version management, reconciliation of automatic vs. human updates, and resilience against changes in the data sources, especially web APIs. It is particularly important to monitor potential decay in the knowledge graph, through the application of existing techniques and methods for decay management coming from scientific domains (see **?** and **?**), where there is a considerable body of work in this area. Once structured as self-contained information packs, personalised subscription, delivery and recommendation of portions of the graph will also be possible.

4 Applying Knowledge Graphs in Cultural Heritage

4.1 Digital Cultural Heritage and Linked Data

Since the first announcement of the Semantic Web vision **?**, there have been a number of existing projects aiming to create open versions of cultural heritage data, including the UK Culture Grid[2] and the Dutch Continuous Access to Cultural Heritage (CATCH) programmes. There are also a number of cultural heritage ontologies in existence, including Categories for the Description of Works of Art (J Paul Getty Trust) and CIDOC CRM **?**. The ontologies and terminologies used are based on a range of technologies, for instance XML and distributed databases as well as RDF/OWL, but there is increasing interest in using techniques from the semantic web in this area. For instance, the OpenART **?** project brings an important arts research dataset, "The London Art World 1660-1735" to the Linked Open Data format so that contents about the art world during that period can be contextualised and linked to the Tate collection and referred to the relevant contemporary art works. In a larger scale, CultureSampo **?** was developed for publishing heterogeneous linked data as a service. The main aim is to create a cultural heritage archive for the whole nation by providing an infrastructures and a set of tools to publish and annotate contents collectively. The case study used in this project is the Finnish cultural heritage archives. The authors argue that semantic linking can add value by facilitating links between artefacts which can lead to better understanding of themes or allow the user to make connections more easily. More recently, the work presented in **?** aims to bring library data into the Linked Data world. The authors discuss some limitations of current data formats in the domain such as MARC[3] and present the process of generating a linked datasets from existing library cataloguing data.

4.2 The Challenges

Despite many efforts to make cultural heritage data open, there have been still several challenges that prevent digital cultural heritage archives from being collected and curated using a bottom-up approach, i.e., directly by community groups. Among these challenges are data heterogeneity and the wide range of computer literacy across the cultural heritage community.

4.2.1 Data Heterogeneity

Data formats between collections and tools for digitizing heritage data are not consistent. For example, some groups may choose to use common multi-purpose tools such as spreadsheets to maintain their archives while others may use pre-existing genealogy software or a relational database. Larger organisations such as national institutions may choose other format to meet their specific requirements such as Key-Value data-stores for performance or RDF triplestores for integration and reusability. Due to the

[2] `http://www.culturegrid.org.uk`

[3] http://www.loc.gov/marc

wide ranging and different types of data formats, it is not trivial for digitalised cultural heritage to be reused and integrated with each other. Therefore, many of such digital archives can only be exploited separately meaning connecting local cultural heritage with national archives cannot be done easily. The integration of knowledge from different digital archives, if possible, has been done only by human. To automate this process, i.e., to integrate/contextualise contents from different archives, it is important to keep data in an open, integrable and reusable format.

4.2.2 Different Levels of Computer Literacy among the Cultural Heritage Community

Because community heritage often being contributed by volunteers and there is a wide range of computer literacy across individuals as well as organisations, it is challenging to provide software platforms that can be used efficiently across individuals and organisations while still allow data reusability and integration. Thus, it is crucial that software and tools supporting community groups in creating and maintaining their cultural heritage data must not involve technical complexities and, at the same time, be familiar by the targeted users so that training and education can be minimised. However, current tools and software for creating Linked Datasets and developing Linked Data applications still require in-depth technical knowledge, which is often not available to the broad community within the cultural heritage domain. A promising approach to tackling this problem is to provide an interface between the Linked Data world (tools and standards) and current tools and platforms that ordinary users are familiar with such as standard Content Management Systems (CMS).

4.3 The CURIOS Project

Two limitations mentioned in the previous section are some of the main motivations behind the CURIOS[4] platform. The aim of the CURIOS project is to enable community groups to preserve and maintain their digital cultural heritage sustainably by combining existing open-source software and open data formats. Firstly, the problem of data integration and reusability can be resolved by using Linked Open Data (e.g., OWL, RDF) as the open data standards. Secondly, to assist individuals and groups with different level of computer literacy to create and maintain their linked datasets, a Linked Data adaptor to existing CMSs were developed.

By combining Linked Data standards and software with Drupal, a popular open-source CMS, CURIOS provides users with limited knowledge on semantic technology a friendly front-end in order to produce linked data (and hence to construct the associated knowledge graphs) without requiring a high level of competency in the underlying technologies (e.g., SPARQL, RDF). In CURIOS, data entered by users are stored in an RDF store while the configuration of how data are presented to users is stored in Drupal's traditional SQL database. This approach allows the linked dataset maintained by CURIOS to be loosely coupled to Drupal so that it can be reused by different applications and re-purposed in different contexts.

Not only supporting the construction of the knowledge graph, the CURIOS system also facilitates services for exploiting the graph such as semantic searching via the use of SPARQL and the semantic database, configurable presentation and visualisation services to the cultural heritage linked datasets. It should be noted that although CURIOS has been deployed in various case studies within the cultural heritage domain, it can still be used as a general-purpose platform which can be applied in other domains.

Within the cultural heritage domain, CURIOS has been used in the following case studies.

[4] Cultural Repositories & InfOrmation System

Hebridean Connections has been the main CURIOS case study and was carried out in collaboration with historical societies based in the Western Isles of Scotland. Previously cultural heritage data about the area had been collected, archived and presented using a proprietary software. However, there were several limitations with this approach that did not allow the collections to be maintained in a sustainable way. Firstly, data entry had to be done via a proprietary software and hence limited the collaborative contribution of volunteers given the uncertain funding sources. Secondly, the archive was kept in a relational database, which make it difficult to be re-used, re-purposed or integrated into other cultural datasets. Based on the original (relational) database schema and suggestions from the historical societies, an ontology for modelling Hebridean Connections archives has been constructed and a subset of this ontology has become the CURIOS upper ontology, the starting point for constructing other knowledge graphs.

Portsoy is another case study about using the CURIOS platform to preserve cultural heritage of Portsoy, a small fishing village on the North East coast of Scotland. In this case study we also worked with the local historical society in constructing the ontology modelling their cultural heritage data. The main difference of this study compared to the Hebridean Connections one is that there was not a database so that we can construct an ontology from, and hence significant effort for knowledge engineering tasks were required.

CURIOS Mobile is an extension of the CURIOS project that explores how cultural heritage linked datasets can be exploited for tourist mobile apps in a rural context **?**. In this project, the Hebridean Connections linked dataset was presented on mobile devices. In addition we investigated on how semantic technology might be used to improve tourists experience in rural areas, where the Internet connection is either missing or unreliable. To overcome the connectivity issues, different caching algorithms based on the knowledge graphs have been proposed. A mechanism to rank the records (URIs) based on the level of interests and a narrative generator from RDF triples were also presented.

POWKist is a follow-up project of CURIOS, which focuses on smaller scale collections such as personal diaries, shoe boxes, etc. The main case study in this project is the diaries of Allan Houston, a prisoner of war (POW), during his time in World War II camps. Unlike the main CURIOS project, POWKist investigates how best to visualise cultural heritage data in an "exhibition" format, meaning that only a selection of the datasets are picked, curated and presented to the viewers. In this project, we also explore how the navigation of linked data-based contents can be enhanced to deliver higher user experience while browsing the collections.

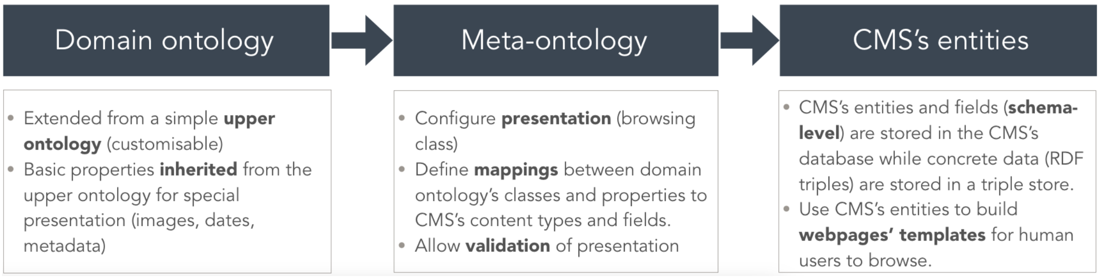
Funeralscapes looks at using CURIOS in a different context, namely for storing research data and supporting academic collaborative work. In this case study, research data about pre-Christian and Viking ancient burial sites, including text- and media-based materials are preserved, linked and presented using the CURIOS system. Audio and video materials are hosted by other services such as `soundcloud.com` and `youtube.com` and presented on a CURIOS website via *oEmbed* formats[5]. This approach brings extra flexibility and scalability into the CURIOS system as popular media stores supporting *oEmbed* (e.g., Instagram, Flickr, Youtube) can be used to host media files, a popular mean to preserve and present cultural heritage data.

4.4 Constructing the Knowledge Graph

The knowledge graphs constructed for these case studies were built on top of an *upper ontology*, which specifies the key classes and properties to be used in the extended ontologies. For example, all record

[5] http://www.oembed.com/

types are a sub-class of `hc:Subject` and have `dc:title` and `dc:description` (used as in the Dublin Core Schema)[6] to specify the title and description of a record. In addition, the upper ontology specifies other special classes and properties such as ones holding metadata or ones used to visualise images and media items.



Domain ontology

- Extended from a simple **upper ontology** (customisable)
- Basic properties **inherited** from the upper ontology for special presentation (images, dates, metadata)

Meta-ontology

- Configure **presentation** (browsing class)
- Define **mappings** between domain ontology's classes and properties to CMS's content types and fields.
- Allow **validation** of presentation

CMS's entities

- CMS's entities and fields (**schema-level**) are stored in the CMS's database while concrete data (RDF triples) are stored in a triple store.
- Use CMS's entities to build **webpages' templates** for human users to browse.

Figure 2: Mapping from a domain ontology to CMS's entities

The whole process of constructing a CURIOS knowledge graph can be summarised in Figure 2. Currently, constructing the domain ontology (i.e., an ontology modelling the archive) is the step that consumes most time and efforts in each CURIOS installation. This is due to the fact that the end-users often do not have sufficient background on ontologies and modelling techniques to design an ontology suits their needs best. To assist users in constructing their own domain ontology, we use parts of the Hebridean Connections ontology as the upper ontology, as mentioned earlier. As long as the domain ontology is produced, a *meta-ontology* specifying mappings between the classes and properties of the domain ontology and the entity types and fields of the CMS are auto-generated. Another important role of the meta-ontology is to allow *validation* of the presentation and data entry. For instance, based on the information of domain/range restrictions in the domain ontology, it is possible to only allow certain types of instances to be linked to each others via a particular object property, e.g., the "child of" relationship can only link a person to another person but not any other types. After having the meta-ontology, the corresponding CMS's entities such as entities types and fields are created, attached and linked to each other.

4.5 CURIOS – A Linked Data Adaptor for Content Management Systems

When a knowledge base is created, it is essential to provide means to produce and consume such knowledge, i.e., a writer and a reader. Instead of re-inventing the wheel, we built a Linked Data adaptor to a popular CMS, namely Drupal. As each content management system will have its own data structures, these data structures can be mapped onto corresponding ontological data structures such as classes or data/object properties. For instance in Drupal, there are entity types and fields which can then be mapped onto class and properties and vice versa. Certainly some configurations specifying which ontology classes are mapped to which entity types are needed. Given a domain ontology, we can generate such configurations automatically (see Section 4.4).

To be able to access linked datasets (in RDF triple format), the database must be a triple store instead of relational databases. This type of back-end databases therefore requires a different query language, namely SPARQL **?**, instead of SQL. Fortunately, a SPARQL query builder is available as a Drupal module,i.e., SPARQL Views **?**, which allows users to generate SPARQL queries in a user-friendly manner.

[6] In practice, the sub-properties of these data properties are included in the extended ontology, such as `hc:title` and `hc:description` in the Hebridean Connections ontology.

However, SPARQL Views requires some background knowledge on Semantic Technology to generate the correct SPARQL queries and also does not provide facilities to update the RDF database.

Our system, CURIOS, uses SPARQL Views as a dependent module to read data from an RDF database, or in other words, to generate SPARQL queries and present results. This has a couple of advantages. Firstly, the CURIOS users do not need to concern about generating correct SPARQL queries, and hence do not need in-depth knowledge on Semantic Technologies (e.g., RDF, SPARQL). Secondly, updates on SPARQL language specification can be dealt with by the developer of the SPARQL Views module instead of CURIOS. To overcome the lack of writer for RDF databases, CURIOS integrates a facility to produce UPDATE queries for SPARQL such as `INSERT`, `DELETE`, etc.[7]

4.6 Presenting and Visualising Cultural Heritage Knowledge Graphs

Constructing knowledge graphs itself only will not make sense without the final step: disseminating knowledge to the wider public. This step is vital in the Cultural Heritage domain as it would attract public awareness, raise funding opportunities and hence increase sustainability of cultural heritage projects, especially for community-based projects. In this section, we discuss how cultural heritage can be made more accessible to the wider public. In particular, we describe the presentation and visualisation layers of the CURIOS system.

Generally, in CURIOS the presentational and modelling layers of the knowledge graphs are separated to take advantages of the well-defined semantics and the adaptive and flexible user interface. For examples date values can be entered in different formats and mapping services of different providers such as Ordnance Survey or Open Street Map can be supported.

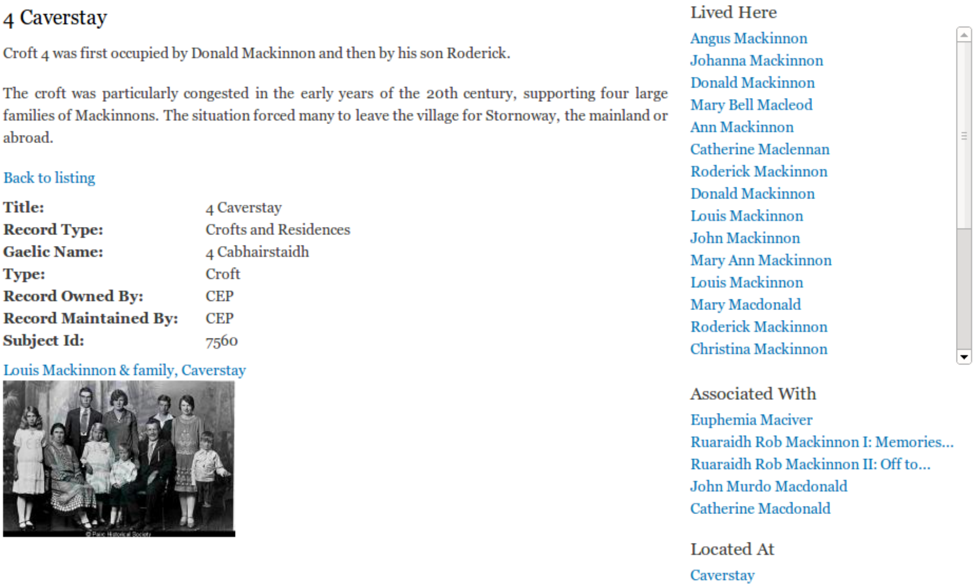
4.6.1 Web-based Presentation of Instances in the Knowledge Graphs

Data held in the CURIOS knowledge graphs are presented as web pages, as discussed in **?**. Each instance in the knowledge graph has a dedicate web page that shows information of that specific item such as title, description, data properties and object properties. Figure 3 demonstrates how an instance of a CURIOS knowledge graph, in this case a residence/croft in the island, is presented. Data properties are summarised in an “Info-box”, object properties are presented in forms of hyperlinks (on the right) to other instances’ web pages. Some special object properties linking the current instance to instances of media types such as images, audio and video are treated differently. These associative media instances are presented as not only hyperlinks but also galleries (for images) or suitable embedded media players (for sound and video items). Aggregated presentation of instances is also supported: for example it is possible to present a collection of the knowledge graph’s instances in a table or on a map, as illustrated in Figure 4a and Figure 4b.

4.6.2 Presenting Data under Inconsistency and Vagueness

A difficult yet unavoidable challenge when using Linked Data (or any dataset with a well-defined semantic) in the Cultural Heritage is *inconsistency* and *vagueness*. The best example would be the case of temporal data. Figure 5 shows some statistics of the date patterns used in the Hebridean Connection corpus **?**. The first column describes the categories of date patterns in the corpus. Column 2 and 3 show the patterns in detail and a representative example of each pattern respectively. Column 4 and 5 show the total count (frequency) of each pattern and the total count of the category. The last column indicates

[7] Note that there is currently no `UPDATE` statement in SPARQL. An update can be represented as a combination of a deletion followed by an insertion.



4 Caverstay

Croft 4 was first occupied by Donald Mackinnon and then by his son Roderick.

The croft was particularly congested in the early years of the 20th century, supporting four large families of Mackinnons. The situation forced many to leave the village for Stornoway, the mainland or abroad.

Back to listing

Title:	4 Caverstay
Record Type:	Crofts and Residences
Gaelic Name:	4 Cabhairstaidh
Type:	Croft
Record Owned By:	CEP
Record Maintained By:	CEP
Subject Id:	7560

Louis Mackinnon & family, Caverstay

Lived Here

- Angus Mackinnon
- Johanna Mackinnon
- Donald Mackinnon
- Mary Bell Macleod
- Ann Mackinnon
- Catherine Maclennan
- Roderick Mackinnon
- Donald Mackinnon
- Louis Mackinnon
- John Mackinnon
- Mary Ann Mackinnon
- Louis Mackinnon
- Mary Macdonald
- Roderick Mackinnon
- Christina Mackinnon

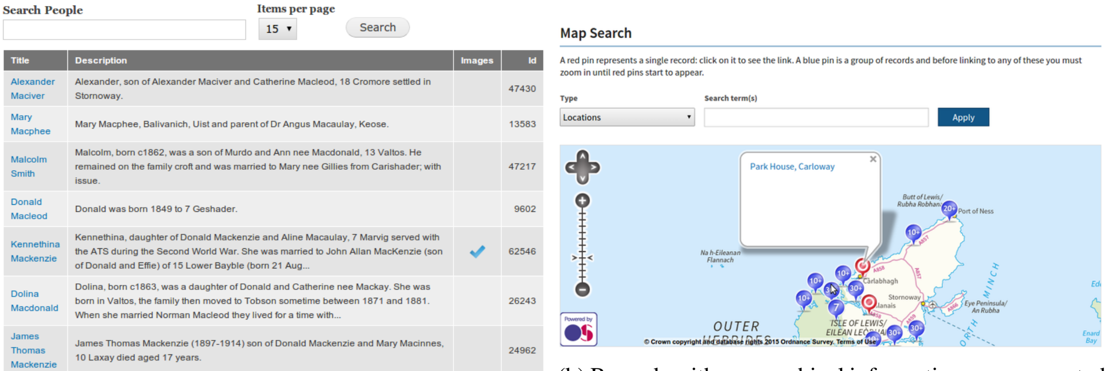
Associated With

- Euphemia Maciver
- Ruaraidh Rob Mackinnon I: Memories...
- Ruaraidh Rob Mackinnon II: Off to...
- John Murdo Macdonald
- Catherine Macdonald

Located At

- Caverstay

Figure 3: Details of a croft **?**



Search People

Items per page 15

Title	Description	Images	Id
Alexander Maciver	Alexander, son of Alexander Maciver and Catherine Macleod, 18 Cromore settled in Stornoway.		47430
Mary Macphee	Mary Macphee, Balivanich, Uist and parent of Dr Angus Macaulay, Keose.		13583
Malcolm Smith	Malcolm, born c1862, was a son of Murdo and Ann nee Macdonald, 13 Valtos. He remained on the family croft and was married to Mary nee Gillies from Carishader; with issue.		47217
Donald Macleod	Donald was born 1849 to 7 Geshader.		9602
Kennethina Mackenzie	Kennethina, daughter of Donald Mackenzie and Aline Macaulay, 7 Marvig served with the ATS during the Second World War. She was married to John Allan MacKenzie (son of Donald and Effie) of 15 Lower Bayble (born 21 Aug...	✓	62546
Dolina Macdonald	Dolina, born c1863, was a daughter of Donald and Catherine nee Mackay. She was born in Valtos, the family then moved to Tobson sometime between 1871 and 1881. When she married Norman Macleod they lived for a time with...		26243
James Thomas Mackenzie	James Thomas Mackenzie (1897-1914) son of Donald Mackenzie and Mary Macinnes, 10 Laxay died aged 17 years.		24962

(a) A list of people records **?**

Map Search

A red pin represents a single record: click on it to see the link. A blue pin is a group of records and before linking to any of these you must zoom in until red pins start to appear.

Type: Locations

Search term(s)

(b) Records with geographical information are aggregated and presented on an Ordnance Survey map

Figure 4: Presenting Linked Data in CURIOS

General Class	Pattern	Example	Frequency	Subtotal	Covered
Exact to the day	y-m-d	1780-06-13	12949		
	d-m-y	10/6/45	725		
	d-M-y	12 MAY 1780	272		
	M-d-y	May 12 1780	8		
				13954	yes
Exact to the month	y-m	1780-12	274		
	M-y	Aug 1780	443		
	m-y	03/1780	2		
				719	yes
Exact to the year	y	1978	10825	10825	yes
Exact to the decade	dec	IN 1860'S	1415	1415	yes
Exact to a range of years	y-y	1939-45	242		
	beforey	pre 1918	2		
	aftery	AFT 1890	3		
				247	yes
Exact to the century	cent	20th Century	4	4	yes
Vague within less than a month	mend	Aug/Sept 1972	26	26	yes (using a date range)
Vague within more than a month but less than a year	yend	1978/79	7	7	yes (using a date range)
Vague year	cy	C. 1932	566		yes (using a date range)
	moddec	early 1950s	86		
				652	
Vague around a decade	cdec	c 1950s	2		yes (using a date range)
	modcent	LATE 1600S	3		
				5	
Not directly interpretable as a date	unk	D.I.I.	3069	3069	no
GRAND TOTAL				30923	

Figure 5: Analysis of Date Forms in the Corpus **?**

whether the pattern can be modelled and presented using the CURIOS system. The first six categories present patterns which are exact to a specific range of time, e.g., exact to a day, a month, a year up to a century. Note that even the range of time is wide (e.g., up to a century timespan), data within these patterns are still interpretable without users' pre-defined semantics. For example, "Aug 1780" can be interpreted as a date within the range from 01/08/1780 to 31/08/1780 inclusively while "May 12 1780" is clearly interpreted as "12/5/1780". The next four categories describe patterns which are vague (i.e., it is impossible to specify a precise range of time) but can still be modelled and presented given a pre-defined interpretation. For instance, if the users or modellers agree on a definition of "Winter YYYY" to be the the last month of the year "YYYY" and the first 2 months of the following year, it is possible to represent "Winter 1780" as a time period from 01/12/1780 to 28/02/1781. The customisable semantics for temporal data brings flexibility to the CURIOS as different user groups will have different interpretations of a single pattern. As an example, users from Australia might have "Winter YYYY" interpreted in a different way compared to users from Scotland due to the difference in geographical contexts.

As can be seen, only 45% of the date representation is exact to the date (the first category) and about

10% (the last category) of date occurrences is uninterpretable as a date. To deal with the remaining cases (making up 45%), we proposed to use the notion of a date range to model dates in the CURIOS system. Some facilities to integrate mappings of date values from the presentational layer (e.g., "Winter 1780") to the modelling level (e.g., a date range from 01/12/1780 to 28/02/1781). However, even when a date value is exact, inconsistency remains a problem as date values are entered into the system in different ways, e.g., some are "y-m-d" while other is "d-m-y". To overcome this, CURIOS provides a user interface for data entry so that dates values can be enter in a consistent format (as a date range). For more details of how inexact dates are treated in CURIOS, we refer readers to **?**.

4.7 Collaborative Construction and Maintenance of Cultural Heritage Knowledge Graphs

As CURIOS was designed as a bottom-up approach to collecting and preserving cultural heritage data, the community groups have been the focus of this project since the beginning. The main difference between community-level user groups and institution-level user groups is that the former heavily relies on the contribution of volunteers in terms of time and efforts to create and maintain data. Therefore, CURIOS provides support for collaborative work in not only the construction and maintenance of the knowledge graph but also the validation of data. This feature is discussed in detail in **?**.

Firstly, we use some metadata data properties in the ontology such as "Approved for publication", "Owned by society", "Maintained by society", "Revision notes" to store the information about publication status and revision logs. One might argue that why the ontology needs to hold such metadata as the facilities for authoring and validating are already available in the CMSs. However, our approach has an advantage of the dataset being loosely-coupled to the tool/service used to maintain it. For example, the knowledge graph can then be easily exported and edited by other tools such as RDF triplestores or Protege[8] in addition to the Drupal CMS. This is very flexible for bulk-update and bulk-import scenarios that often happen in the Cultural Heritage domain.

Secondly, the CURIOS system also employs the roles and permissions feature in the Drupal CMS to design a *flexible user permission scheme* that can be tailored to adapt the organisational structure and policies of different user groups. For example, in the Hebridean Connections case study, permissions of a data creator and a data validator must be mutually exclusive, i.e., a creator cannot validate and publish a record which she has just created. Moreover, a member of a historical society cannot edit or validate a record created by member of another society. In another case study, Funeralscapes, there is only one group of editors who can create, edit and validate any records.

5 Conclusion

In this deliverable, we continued the D9.1, in which we have considered one of the key Knowledge Graph technique - Named Entity Resolution and we have defined a Diagnostics Framework for troubleshooting and optimizing corresponding systems in industrial scenarios. Our motivation for this work has been the empirical fact that a NER system's satisfactory performance in a given scenario does not constitute a trustworthy predictor of its performance in different settings. As industrial clients typically expect a high and consistent performance from the NER solutions they pay for, our framework helps NER system developers and consultants identify the reasons why their system performs unsatisfactorily in a given scenario and take appropriate actions to increase performance.

[8] http://protege.stanford.edu

In defining our Knowledge Graph we have first identified the main factors that affect NER effectiveness; two of these are i) the level of ambiguity that characterizes the scenario's entities and ii) the adequacy of the contextual evidence applied for disambiguation. Then we have defined metrics and processes for quantifying these factors and we have linked the values of these metrics to specific actions. In this deliverable we describe two cases where the application of the diagnostic framework helped us to significantly increase the (initially low) effectiveness of Knowledge Tagger, our in-house developed NER system. Knowledge Tagger uses primarily ontological knowledge graphs as disambiguation evidence.

In D9.2, we have present success stores of the applications of knowledge Graph techniques from various domains (healthcare, media and culture) and different organisations (international company - IBM, Small and Medium Enterprise - HAVAS, and University - the University of Aberdeen).

Acknowledgement

This research has been funded by the European Commission within the 7th Framework Programme/-Marie Curie Industry-Academia Partnerships and Pathways schema/PEOPLE Work Programme 2011 project K-Drive number 286348 (cf. `http://www.kdrive-project.eu`).