Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Biomedical Informatics: X 6–7 (2020) 100074

Contents lists available at ScienceDirect

Journal of Biomedical Informatics: X

journal homepage: www.journals.elsevier.com/journal-of-biomedical-informatics-x

Interoperability of population-based patient registries

Nicholas Nicholson*, Andrea Perego

European Commission Joint Research Centre, Italy

ARTICLE INFO

Keywords:
Population-based patient registries
Interoperability
Federated semantic metadata registry framework
ISO/IEC 11179
Linked Open Data

ABSTRACT

Enabling full interoperability within and between population-based patient-registry domains would open up access to a rich and unique source of health data for secondary data usage. Previous attempts to tackle patient-registry interoperability have met with varying degrees of success, but a unifying solution remains elusive. The purpose of this paper is to show by practical example how a solution is attainable via the implementation of an existing framework based of the concept of federated, semantic metadata registries. One important feature motivating the use of this framework is that it can be implemented gradually and independently within each patient-registry domain. By employing linked open data principles, the framework extends the ISO/IEC 11179 standard to provide both syntactic and semantic interoperability of data elements with the means of specifying automated extraction scripts for retrieval of data from different registry content models. The examples provided address the domain of European population-based cancer registries to demonstrate the feasibility of the approach. One of the examples shows how quick gains are derivable by allowing retrieval of aggregated core data sets. The other examples show how aggregated full sets of data and record-level data might also be retrieved from each local registry. An infrastructure of patient-registry domains adhering to the principles of the framework would provide the semantic contexts and inter-linkage of data necessary for automated search and retrieval of registry data. It would thereby also lay the foundation for making registry data serviceable to artificial intelligence (AI) applications.

1. Introduction

Whereas no consistent definition exists for the term patient registry – possibly underlying the many different purposes for which they are used [1–3] – an important qualifier is attached to the definition of population-based registries [4].

Within the domain of cancer, the principal aim of population-based cancer registries is to record all new cancer cases arising in a defined population with emphasis on epidemiology and public health practice [5]. Population-based cancer registries provide information on the cancer burden for healthcare planning and evaluation purposes, and also provide valuable data for studies on prevention, early detection/screening, and cancer-related healthcare. As an example, it is of general public interest to know the risk (and its evolution over time) of developing or dying from a particular cancer. Such information is obtained by epidemiologists and used by public health planners to effect changes in healthcare practice [6].

The concepts of population-based cancer registries are equally applicable to other disease paradigms and we therefore introduce the encompassing term: Population-Based Patient Registries (PBPR).

PBPRs attempt to capture all cases related to a specific illness/condition within a defined population, which is important for removing sources of selection bias from epidemiological studies. PBPRs are therefore instrumental in the planning and evaluation of disease control programmes as well as in the effectiveness of patient healthcare measures.

Data collection in a PBPR is a time-consuming and labour-intensive undertaking requiring access to a number of different data sources that include hospital discharge records, clinical records, pathology reports, and death certificates, all of which may use different encoding schemes for disease. Painstaking commitment is required to ensure the quality of the registry's data. Data quality can be compromised through such things as undiagnosed cases, uncertainty of diagnosis, under-reporting of cases, and inaccurate application of codes [7]. PBPRs are different from clinical registries, where the focus is on clinical care and hospital administration. PBPRs collect fewer variables and the variables they do collect are at less granular detail for purposes of comparison at population levels. Consequently PBPRs and clinical registries are used for quite different purposes, which serves to explain the conflicts that can arise between clinical demands for prognostic precision and epidemiological demands for comparability and completeness [8].

* Corresponding author.
E-mail address: nicholas.nicholson@ec.europa.eu (N. Nicholson).

https://doi.org/10.1016/j.yjbinx.2020.100074
Received 18 February 2020; Received in revised form 29 April 2020; Accepted 14 May 2020
Available online 13 June 2020

Abbreviation

BBMRI-ERIC *Biobanking and Biomolecular Resources European Research Infrastructure*
CCD *Continuity of Care Document* – A document exchange standard for sharing patient summary information between computer applications
C-CDA *Consolidated CDA* – A structured data format for standardising content and structure for clinical care summaries
CDA *Clinical Document Architecture* – XML-based, electronic standard used for clinical document exchange
CDE *Common Data Element* – A standard format for data elements
COEUS A semantic web application framework
CR *Cancer Registry*
CS *Classification Schemes* – Part of the classification metamodel of ISO/IEC11179 intended to permit the classification of arbitrary objects into hierarchies
CSI *Classification Scheme Item* - a metadata item that might be classified in a classification scheme
CTV-3 *Clinical Terminology Version 3* – later version of READ codes version 2. Used in the UK's National Health Service
DEC *Data Element Concept* – Used in ISO/IEC 11179 to denote the association of Object Class and Property
DICOM *Digital Imaging and Communications in Medicine* – International standard to transmit, store, process, and display medical imaging information
ECIS *European Cancer Information System*
ECFSPR *European Cystic Fibrosis Society Patient Registry*
EHR *Electronic Health Record*
ENCR *European Network of Cancer Registries*
EU *European Union*
EUBIROD *European best information through regional outcomes in diabetes*
EUROCISS *European Cardiovascular Indicators Surveillance Set*
FAIR *Findable, Accessible, Interoperable, Reusable* – Guiding principles for making data available
FHIR *Fast Healthcare Interoperability Resources* – Next-generation standard of HL7 for electronic exchange of healthcare information
GDPR *General Data Protection Regulation of the European Union*
HITSP *Health Information Technology Standards Panel*
HITSP C32 The HITSP Summary Documents Using HL7 Continuity of Care Document (CCD)
HL7 *Health Level Seven* – A not-for-profit standards-developing organisation in the field of electronic health information
HTTP *Hypertext Transfer Protocol* – application-layer protocol for transmitting hypermedia documents
ICD *International Classification of Diseases*
ICD-10 *ICD 10th revision*
ICD-O-3 *International Classification of Diseases for Oncology, 3rd edition* – classification system of tumours according tumour topography, morphology, behaviour, and grade
IHE *Integrating the Healthcare Enterprise*
IHE DEX *Data Element Exchange* – an integration profile in the IHE QRPH domain
IHE ITI *IT Infrastructure* – one of IHE's technical framework domains
IHE PCC *Patient Care Coordination* – one of IHE's technical framework domains
IHE QRPH *Quality, Research, and Public Health* – one of IHE's technical framework domains
IHE RFD *Retrieve Form for Data Capture* – an integration profile in the IHE ITI domain
IHE XDS *Cross-Enterprise Document Sharing* – an integration profile in the IHE ITI domain
ISO/IEC *International Organization for Standardisation/International Electrotechnical Commission*
ISO/IEC 11179 *ISO/IEC Metadata Registry Standard*
LOD *Linked Open Data* – Structured open data that are linked to other data in machine-readable ways
MDR *Metadata Registry* – A registry containing metadata rather than actual data
METeOR *Metadata Online Registry* – Australia's Metadata Online Repository for national metadata standards for the health, aged care, community services, early childhood and housing and homelessness sectors
MIABIS *Minimum Information About BIobank data Sharing*
OC *Object Class* – One of the component parts of a CDE as defined by ISO/IEC 11179, which together with the Property component constitutes a DEC
OWL *Web Ontology Language* – family of computer languages with different degrees of description logic expressivity for creating web-based ontologies/knowledge bases
P *Property* – One of the component parts of a CDE as defined by ISO/IEC 11179, which together with the Object Class component constitutes a DEC
PARENT *Cross Border PAtient REgistries iNiTiative*
PBPR *Population-Based Patient Registry* – A registry that collects a set of patient-related data from a defined population base
RD-Connect Integrated platform connecting databases, registries, biobanks, and clinical bioinformatics for rare disease research
RDF *Resource Description Framework* – A standard model for data interchange on the Web
READ codes Clinical terminology system used in the United Kingdom's National Health Service until superseded by SNOMED CT
REST *Representational State Transfer* – Architectural style that builds on HTTP for developing web services
RESTful Conforming to the REST architectural style
RoR *Registry of Registries* – A registry that holds information on registries
SCTID *SNOMED CT Identifier*
SKOS *Simple Knowledge Organization System* – A specification for supporting knowledge organisation systems on the web such as thesauri, taxonomies, and classification schemes
SNOMED *Systematized Nomenclature of Medicine* – A comprehensive, multilingual clinical healthcare terminology with an ontological foundation based on OWL
SNOMED CT *SNOMED Clinical Terms*
SNOMED RT *SNOMED Reference Terminology* – Precursor to SNOMED CT, which was merged with CTV-3 to form SNOMED CT
SNOP *Systematized Nomenclature of Pathology* – Precursor to SNOMED
SPARQL *Recursive acronym for SPARQL Protocol and RDF Query Language* – A computer language used for querying and manipulating data stored in RDF format
SQL *Structured Query Language* – A computer language used for querying and editing databases
TNM *Tumour, Nodes, Metastasis* – Globally recognised standard for classification of malignant tumours
Turtle *Terse RDF Triple Language* – One of the syntax and file formats for expressing data in RDF
UMLS *Unified Medical Language System* – A unified vocabulary for mapping concepts in multiple vocabularies
URI *Uniform Resource Identifier* – A unique string-type identifier of a web-based resource
VD *Value Domain* – One of the component parts of a CDE as

	defined by ISO/IEC 11179. A DEC can associate with any number of VDs depending on the representation of the data type
WHO	*World Health Organization*
XML	*Extensible Markup Language*
XPath	*XML Path Language* – A language for navigating through and searching XML documents
XSD	*XML Schema Definition* – Defines the elements and attributes in an XML document

1.1. Objective

Encouraging secondary data usage of PBPRs would further the symbiosis between data usefulness and data quality. It has been observed that using population-based registry data not only reduces the time and costs otherwise spent on epidemiological studies but leads to increased validity of results [9]. Moreover, linkage of registry data with other types of data, such as environmental, socioeconomic or dietary/lifestyle data covering the same populations can stimulate more specific and targeted research to test the validity of any observed correlations.

In this paper we present a means based on an existing metadata framework by which this linkage could be achieved at a technical level both for aggregated data sets and individual record level data. Most of the examples are with reference to the chronic diseases domain but PBPRs are vital also for the infectious diseases domain as has so clearly been underlined by the recent Covid-19 coronavirus pandemic.

2. Standards for healthcare data interoperability

Attempts to link health data are soon frustrated by the need to align different systems used for the various operations of collecting, recording, describing, and classifying information. Nowhere is this more apparent than in the area of electronic health records (EHR).

Considerable effort has been expended over many years in the drive towards data standards that allow interoperability between disparate EHR systems. Data standards are needed at many different levels, including address protocols, message formats, document architecture, management of document sharing processes, and healthcare terminology [10].

2.1. Message format standards

Two examples of widely used message format standards include Digital Imaging and Communications in Medicine (DICOM) and Health Level Seven version 2 (HL7 v2). As well as facilitating interoperability by ensuring common encoding specifications, they also provide transport-packaging mechanisms for documents conforming to document architecture standards, such as HL7 Clinical Document Architecture (CDA).

2.2. Document architecture standards

Integrating the Healthcare Enterprise (IHE) is an initiative between healthcare professionals and industry to improve healthcare information sharing. Part of this work involves defining integration profiles that provide precise definitions of how standards can be implemented to meet specific clinical needs [11]. The profiles are categorised under domains specific to their clinical and operational scope. Examples of these domains are Patient Care Coordination (IHE PCC), IT Infrastructure (IHE ITI), and Quality, Research, and Public Health (IHE QRPH).

The clinical document architecture standard HL7 CDA provides a hierarchical set of specifications for the structure of clinical documents (essentially electronic versions of physical documents). The Consolidated CDA (C-CDA) implementation guide [12] contains a library of CDA templates, incorporating and harmonizing previous efforts from Health Level Seven (HL7), IHE, and Health Information Technology Standards Panel (HITSP). It represents harmonization of the HL7 Health Story guides, HITSP C32, related components of IHE PCC, and Continuity of Care Document (CCD).

Fast Healthcare Interoperability Resources (FHIR) is a recent, next-generation standard framework created by HL7. FHIR broadly fits into both the categories described under Sections 2.1 and 2.2. It is based on modular components or resources that can be combined to provide customisable solutions for providing clinical and administrative information. It also describes an application programming interface. FIHR combines elements of HL7 v2, HL7 v3 and CDA; FHIR resources provide direct implementation of functionality from other standards [13] including DICOM and IHE. One advantage of FHIR concerns the fact that resources can easily be assembled into working systems at a fraction of the price of existing alternatives and is suitable for use in a wide variety of contexts. Included in these contexts are mobile phone apps, cloud communications, EHR-based data sharing, and server communication in large institutional healthcare providers [14].

2.3. Document sharing specifications

IHE Cross-Enterprise Document Sharing (XDS) is a standards-based specification under the IHE ITI domain for managing the sharing of documents between healthcare enterprises using federated document repositories and a document registry (for indexing and storage document metadata). Since XDS is document content neutral, it supports any type of clinical information without regard to content and representation (e.g. HL7 CDA, DICOM, etc.). Enterprises however need to belong to an XDS Affinity Domain and to agree a common set of policies [15].

Concerning interoperability between EHRs and registries, the IHE profile Retrieve Form for Data Capture (RFD) allows a generic way for systems to interact. Once an EHR is RFD-enabled, it can be used for multiple use-cases by allowing information exchanges for different purposes [1].

2.4. Terminology standards

Examples of terminology standards include the International Statistical Classification of Diseases (in various revisions: ICD-10, ICD-11) and Systematized Nomenclature of Medicine (SNOMED). ICD is a medical classification list maintained by the World Health Organisation (WHO) and defines the universe of diseases, disorders, injuries, and other related health conditions in a comprehensive, hierarchical fashion [16]. SNOMED evolved from the Systematized Nomenclature of Pathology (SNOP) for describing morphology and anatomy. A logic-based version on SNOMED (SNOMED RT) was combined with Clinical Terms Version 3 (CTV-3), which itself evolved from the Read codes to create SNOMED CT [17]. SNOMED CT is designed for a large number of different applications and is especially useful for clinical decision support whereas ICD is more suited to classification.

The Unified Medical Language System (UMLS) is a unified vocabulary that brings together concepts listed in multiple vocabularies. It integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records [18]. It provides a useful tool for mapping between ICD-O-3 codes (International Classification of Diseases for Oncology, 3rd edition) and SNOMED CT terms [19].

2.5. Healthcare data standards in relation to PBPRs

Most of the focus regarding data-interchange standards has been on electronic health records and the need to access health data related to a particular patient or group of patients. While this work is of direct importance to PBPRs in the collection and submission of data to registries, less attention has been paid to the exchange of aggregated data sets that are important for epidemiological studies. A survey of European PBPRs indicated that whereas respondents were most familiar with the HL7 standards, they were not necessarily using them beyond collection of primary data for the reason that the standards were not appropriate to their specific data structure and needed information [20].

Epidemiology is not so much concerned with accessing particular individual cases as it is with selecting complete sets of patient cases sharing certain commonalities from a known population. In this regard, aggregation of cases from a PBPR, or several PBPRs, is particularly pertinent.

3. Examples of European PBPRs and their data sets

PBPRs collect certain information on individual patients in a defined population who have been diagnosed with a given condition. Apart from general information such as date of birth, date of diagnosis, sex, the data variables collected by PBPRs are dependent upon the disease domain and can vary between registries depending on national or regional health policies as well as the resources available to the registry. The most important variables in a specific healthcare domain form what is called the common or core data set. The variables in the core data set are generally the most harmonised since they are compared across regional and national boundaries.

Examples of core data sets associated with European-based PBPRs in a selection of healthcare domains are: cystic fibrosis [21], cardiovascular disease [22], congenital anomalies [23], diabetes [24], rare diseases [25], and cancer [26]. As a specific example, the variables in the European population-based cancer-registry core data set capture information concerning the tumour such as: topography (tumour location); morphology (tumour form/structure); behaviour (whether the tumour is benign/in situ/malignant/uncertain); grade (the degree of the abnormality of the tumour cells); basis of diagnosis (how the tumour was diagnosed); and stage (the state of progression of the tumour at diagnosis). The latter is generally described by the TNM Classification of Malignant Tumours globally recognised standard [27].

Whereas PBPRs hold specific information on patients in defined populations, the focus of their work is not so much at the individual level. Epidemiology requires individuals' information in order to identify the relevant cohorts of patients for a particular study. Once the cohorts have been identified, the personal identifiers are removed and results are provided as aggregated data.

Data may be aggregated in a number of ways. One example is by age group whereby number of cases (incidence, mortality, etc.) is aggregated in predefined age ranges. In case of rare occurrences of a specific type of disease, where the number of cases is low, data may also need to be aggregated across wider geographical areas to avoid potential identification of individuals – this is particularly the situation encountered with rare-disease registries (RDRs).

Indicators (such as incidence, mortality, survival, and prevalence) derived from the core data set provide the means of comparing the disease burden between different populations. Considerable effort is expended in ensuring the accuracy and comparability of the indicators and underlying data and therefore it is important to allow maximum reuse where possible.

4. Data interoperability needs of PBPRs

It is oftentimes not a straightforward matter even to find the core data sets. In addition, the descriptions of the associated variables are not necessarily defined in rigorous and unambiguous terms. Just addressing these two aspects alone would bring an immediate breakthrough in the possibility of mapping heterogeneous data sets along common fields of aggregation.

In order to appreciate more fully the data interoperability needs for effective secondary use of PBPRs, they can be considered as a number of distinct use cases:

4.1. Use case #1 – Access to aggregated core data sets

The first use case relates to the need of collecting harmonised data (corresponding to the core data sets discussed in Section 3) from a number of PBPRs. This use case is an example of processes already in operation to gather datasets for comparison at European Union (EU) level for monitoring the burden of disease in different healthcare domains. Currently this is undertaken via some central entity issuing a call for data to the participating registries. The call specifies a common data protocol to which registries are expected to adhere in submitting their data. The collection entity validates the data against a harmonised data validation protocol, which may require a number of iterations in the data submission. Once all the data sets have been validated, the data are aggregated and made publicly available. The process is not optimal for three main reasons. Firstly, it introduces significant time delays on top of those already incurred by the registries themselves in collecting data from the primary sources; when the aggregated data sets are finally made available they may be several years out of date. These delays compromise the value of the data for timely feedback into healthcare planning processes. Secondly, it is demanding of resources – the iterations are relatively manual and require a significant number of communication workflows. Thirdly, data sets are thereby duplicated, leading eventually to data integrity and versioning issues.

4.2. Use case #2 – Access to aggregated full data sets

The second use case is an amplification of the first use case but would allow maximum reuse of the registry data by access to the registry's aggregated full data set. Access to the core data set has been described in Section 4.1, but this forms only a subset of the data stored by the registry. Whereas variables outside the core data set may not be standardised or harmonised, if they were described following standard metadata concepts their meaning would be clearer for data users to analyse them in an appropriate way with less danger of making false assumptions. Indeed a rich source of untapped data resides in the variables of the full data sets. This use case therefore introduces the notion of a registry's aggregated full data set.

4.3. Use case #3 – Access to record level data

The third use case concerns the situation in which access is needed to individual record level data. Such a scenario may be faced in research studies that need to select a set of case records across several registries and then perform the aggregation. This use case is trickier regarding the data privacy requirements since registries would in this case need to release individual record-level data, albeit to another registry. This use case is primarily needed in the case of rare diseases where one registry may not have a sufficient number of cases to undertake a particular study or for high-resolution studies in the absence of solutions for use case #4 described in Section 4.4.

4.4. Use case #4 – Aggregation on demand

The fourth use case concerns a registry service that could be termed aggregation on demand. As discussed in Section 3, the focus of epidemiology is not the individual per se, but groups of individuals sharing a common condition. The normal procedure follows a request to registries interested in participating in a given research project.

Depending on data privacy agreements, the study proceeds with individual case records and the results are published in terms of aggregated data with no reference to specific individuals. If instead a means were available of aggregating data according to a study-dependent aggregation protocol, there may be no need for registries and studies to set up data-protection agreements and protocols that add time delays and costs to projects. This use case would require a standard way of specifying aggregation protocols that could be simply applied to registry data.

4.5. Use case #5 – trace-back to primary data sources

The fifth use case also concerns high-resolution studies. Analysis of aggregated data sets may suggest correlations or patterns, the statistical validity of which may require investigation in greater detail via high-resolution studies. These studies generally require more specific data and it is therefore necessary to identify the individual data subjects of those constituting the aggregated set of interest. It is not a straightforward process to trace back this information to the primary data sources and the exercise is a costly process both in time and resources. Having the possibility to trace back automatically the original primary data source given a (pseudonymised) patient identifier held in the registry would greatly facilitate these sorts of studies.

4.6. Current limitations regarding secondary data utilisation of PBR data

The current limitations regarding access to patient registry data – even within a given patient domain – are widely apparent [3,7] and constitute a first major challenge without regard to the more complex one of linking data between heterogeneous registries covering different patient domains.

The difficulty of understanding which PBPR data sources are available and what sorts of data they hold, coupled with the cost in both time and resources of making that data available in the format required greatly compromises the secondary-data usage of PBPRs. Even if the data were readily available, it is not a straightforward matter for researchers to know how the associated variables relate to the research study in mind or even how the data can be used appropriately. The registry is normally actively involved in the study to ensure the data are used correctly, but this also serves to limit the number of concurrent studies.

Previous attempts have been made at EU level to address some of the underlying needs [28], particularly within the field of rare diseases in which the problem of interoperability is more acute on account of the widely different types of diseases classified within the same overall patient domain [29]. These efforts, however, remain largely focused within each specific patient-registry domain and although the ensuing solutions may ease access to European-harmonised data on a thematic disease level, the use of different metadata methodologies coupled with different data-registration and data-discovery mechanisms will still present a challenge for the inter-linkage of data between registry domains.

Without any overarching strategy and overall coordination across patient registry domains, much effort will continue to be duplicated in terms of reinvention of solutions that have at their basis shared and common requirements. It would be worthwhile to find some way of uniting these efforts towards a common and scalable methodology. With a common framework initiatives would converge more rapidly towards greater data interoperability with consequently greater scope for secondary data usage.

4.7. Requirements of a framework to fit the process constraints

In view of the fact that the data sets may stretch back many years, it is not feasible to require fundamental changes to already existing data representations or individual registry infrastructures. A more practical solution would be one that encouraged, wherever possible, mapping of local registry data structures to common metadata constructs and the means of retrieving data on the basis of those mappings. Given the number of entities involved and the autonomy of those entities, any solution should as far as possible also be standards-agnostic; it would not be realistic to advocate the use of any one standard and the framework should ideally be able to work with the different standards in place.

A framework with the potential of meeting many of these requirements already exists and its capacity of interfacing with different healthcare standards in the field of electronic health records has been demonstrated [30]. The framework consists of a federated architecture of semantic ISO/IEC 11179 [31] metadata registries (MDRs) and provides an innovative way for accomplishing mapping of metadata across systems and semantic contexts using linked open data (LOD) principles.

The semantic MDR framework was initially proposed for enabling data inter- linkage between different EHR formats with a primary focus on mapping the individual common data elements (CDEs) to standard CDE models. The framework has also been successfully applied to secondary use of EHRs for post-marketing surveillance [32]. Combining the concepts of metadata description of ISO/IEC 11179 with the capacities of linked open data and semantic web technologies provides a powerful and highly adaptable data-interoperability framework. The framework essentially extends the concepts behind the IHE Data Element Exchange (DEX) profile [33] under the IHE QRPH domain by federating the metadata registry and providing semantic linkage. It thereby scales the DEX concept across systems of disparate CDE definitions [30].

The versatility of the model would make it amenable to any application requiring standardised data exchange and, arguably, the limitation of its applicability is constrained more by implementation-based decisions within the given domain than by the technological constraints themselves. This paper presents a proposal for an implementation of the framework to address the particular data interoperability issues in the field of PBPRs.

5. ISO/IEC 11179

ISO/IEC 11179 [31] is a general description framework for data of any kind. It was prompted, amongst other things, by the need for standardised data-design procedures in order to ensure the emergence of data elements capable of supporting electronic data interchange. According to the standard, metadata are viewed as data about data in some context, where context can be considered as the set of circumstances, purposes, or perspectives relating to the data elements. In terms of the ISO/IEC 11179, a data element has therefore both semantic and representational components.

ISO/IEC 11179 distinguishes between the concept of a data element (Data Element Concept – DEC, describing the contextual semantics) and its representation (describing the permitted values a data element may use); the set of permitted values is called the Value Domain (VD). The DEC further comprises two sub-components; the Object Class (OC), encapsulating the general underlying concept of the DEC; and the Property (P), a characteristic shared by all members of the OC (independent of any specific data-type). A data element is created when a specific DEC and a specific representation are associated together.

The OC, Property, and VD provide searchable interfaces within the ISO/IEC metadata registry framework. For example, searching on a given Property would provide references to all DEC classes related to that Property and, through these, references are returned to all the associated OCs and VDs. An on-line application of the Australian Institute of Health and Welfare, METeOR [34], provides a good example of the means of searching for metadata elements within an ISO/IEC 11179 metadata registry.

5.1. Implementation example – METeOR

By way of a specific example, METeOR defines one of its OCs as “Person with cancer”. Associated with this OC are a number of Properties, one of which is “Primary site of cancer”. The associated DEC

is “Person with cancer – primary site of cancer” which encapsulates the concept of a person with cancer having a primary site of the tumour.

Since there are various coding schemes for the conceptual domain denoting “tumour type”, the “Person with cancer – primary site of cancer” DEC is free to associate with any of them (via assignment of different VDs), with each association providing a separate data element. Thus one data element is defined as: “Person with cancer – primary site of cancer, code (ICD-10-AM 7th edn) ANN{.N[N]}” for descriptions of tumours according to the 7th edition of ICD-10 (International Classification of Diseases, 10th revision) [35], whereas another data element is defined as: “Person with cancer – primary site of cancer, code (ICD-O-3) ANN{.N[N]}” for descriptions of tumours according to ICD-O-3 (International Classification of Diseases for Cancer, 3rd edition) [36]. In METeOR, the fields “ANN{.N[N]}” refer to the format of the codes (one alphabetical character followed by two numeric characters with an optional decimal point followed by one or two numeric characters). The fact that the codes for these two schemes for classifying cancer-type are expressed in exactly the same format underlines the need to make unambiguous distinction between the VDs, as supported by the ISO/IEC 11179.

5.2. Classification schemes

A further important principle of ISO/IEC 11179 concerns the use of classification schemes, which provide the means for developing metadata with enhanced semantic descriptions. OCs, Properties, VDs, DECs, and Data Elements are all classifiable components, and this aspect is an integral part of the philosophy underlying the federated semantic MDR framework.

By way of illustration, Fig. 1 depicts the constituent concepts of the CDE “Person with cancer – primary site of cancer, code (ICD-O-3) ANN{.N[N]}”.

6. Federated semantic MDR framework

The federated semantic MDR framework enhances the capacity of the ISO/IEC 11179 metamodel by mapping its various constructs to associated OWL (Web Ontology Language) [37] classes and properties within an ontological representation of the metamodel [30]. This mapping allows the different concepts of an ISO/IEC 1179 data element to be described in RDF (Resource Description Framework) [38] and therefore uniquely identifiable and openly accessible through URIs.

6.1. Restful services

As part of the functionality of the framework, semantic MDRs are expected to open some simple REST (REpresentational State Transfer [39]) services. REST is an architectural style that builds on HTTP for developing web services. Web services that conform to the REST architectural style are referred to as RESTful web services. REST uses URIs to identify resources and RESTful web services are able to handle web resources in a variety of different formats.

The set of RESTful web services provided by a semantic MDR [30] are: CDE endpoint (for retrieving the RDF description of a CDE or its components); CDE search (for performing searches on a CDE or its components; Semantic links (for retrieving all the semantic links associated with a CDE); Extraction specification (to retrieve the extraction specification defined for a CDE); and SPARQL [40] endpoint (to provide native SPARQL support).

Dereferencing a CDE or any of its components via the semantic MDR’s RESTful services returns the description of the component and all its context in RDF, including any classification scheme resources. Classification Schemes (CS) and their sub components Classification Scheme Items (CSI) are the means in ISO/IEC 11179 for developing metadata with enhanced semantic descriptions. The semantic MDR uses these resources to annotate the components of a data element with hyperlinks to other MDRs or to terminology systems thereby providing a rich means of searching and linking CDEs or their sub components across domains through LOD principles.

CDEs are generally abstract data element definitions but the underlying data that have been described by these CDEs can be retrieved through the framework. The way in which this achieved is by setting up an extraction-specification CS with an extraction-specification script that is executed on a data server. The semantic MDR framework supports three types of extraction specifications: XPath, SPARQL, and SQL. The mechanism is described more fully in [30] and examples are provided in Section 7.

6.2. Semantic linkage with a terminology system

As an example of linkage with a terminology system, Fig. 2 shows the CDE Property “Primary site of cancer” discussed in Section 5.1 that is associated with a CS resource described by SNOMED CT, which is hosted on the BioPortal web site [41]. Contained within this resource is a CSI resource of type SKOS:exactMatch with value given by the URI to the SNOMED CT code: 399687005 (primary tumour site). Dereferencing the Property within the semantic MDR framework would provide the semantic link to the specific match in SNOMED and provide the means to find all other components in the framework referencing this same code.

The ingenuity of the framework leads to a number of advantages; namely, that:

- it can be implemented gradually in a well-staged approach;
- it requires no fundamental change to the local data – the underlying principle is to map local metadata to standard metadata descriptions or common dictionaries without enforcing compliance of metadata to any particular standard;
- it is scalable across many different patient-registry domains and can moreover be implemented in each domain independently of the other domains;
- not only are CDEs automatically registered and therefore findable, but they can be reused in an interoperable way via the semantic mapping descriptions to harmonised metadata standards.

Furthermore, via these mappings and their associated extraction specifications, local data elements described by otherwise non-standard CDEs are readily accessible. The semantic MDR framework therefore intrinsically supports all four of the FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles for scientific data and stewardship [42].

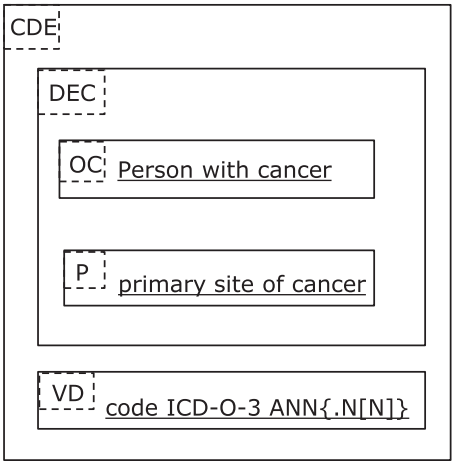

Fig. 1. An example showing the construction of the CDE: “Person with cancer – primary site of cancer, code (ICD-O-3) ANN{.N[N]}” from its constituent concepts according to ISO/IEC 11179, which include the Object Class (OC): “Person with cancer”; Property (P): “primary site of cancer”; and Value Domain (VD): “code (ICD-O-3) ANN{.N[N]}”.

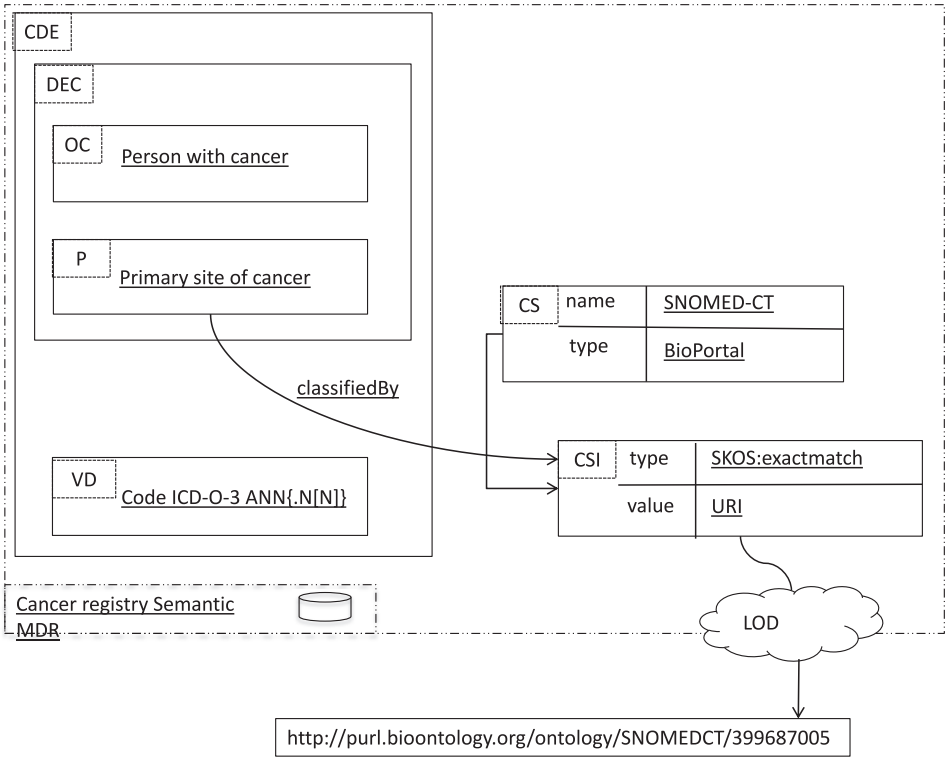

Fig. 2. An example of linkage of a CDE Property with SNOMED CT. The Property "Primary site of cancer" has a CS resource described by SNOMED CT, hosted on BioPortal. Contained in the CS, a CSI resource of type SKOS:exactMatch with value given by the URI is linked to the SCTID code: 399687005 (primary tumour site).

In order to illustrate how the semantic metadata framework can be implemented to address the use cases identified in Section 4, the practical example of population-based cancer registries will be considered.

7. Application of the semantic MDR framework to population-based cancer registries

7.1. Use case #1: Access to aggregated core data sets

The aggregated core data set of the European Network of Cancer Registries (ENCR) consists of five main variables: indicator type (incidence/mortality), sex at birth, cancer site, historical year, and number of cases broken down into five-year age ranges. Within a cancer registry's (CR) tabularised aggregated data set, these variables equate to the column names (c.f. Fig. 3).

Currently these data are accessible from the European Cancer Information System (ECIS) [43], but it requires some effort to extract them in their entirety. Furthermore, the metadata are described in different places [26,44], and not in machine-readable terms, nor do the majority of variables link to more generic metadata terms thereby rendering cross-linkage of data difficult between different PBPR domains.

7.1.1. Transcription of data into RDF

In order to make the aggregated data set accessible via the semantic MDR framework, it is first of all transcribed in RDF, and maintained in a triple store. RDF is used since it provides a convenient means for the data to be searched and accessed via a SPARQL end point.

A triple store stores data according to triplets corresponding to subject, predicate, and object. RDF provides a standard model for data interchange on the Web and allows structured and semi-structured data to be shared across different applications. To transcribe the CR aggregated data in terms of RDF, the column names translate to the predicates of the RDF triples, with the row identifier (or primary key) forming the subject and the data values of the column names forming the objects.

Fig. 3 shows the header and one row of the aggregated data set corresponding to one CR (information taken from ECIS). It shows, within that particular CR, the number of incident male lung cancer cases in the associated age brackets per 100,000 head of male population in 2009. Fig. 4 shows how one of the age-bracket elements (age bracket 55–59) from this row of data would be transcribed in RDF. The predicates and the objects – apart from the xsd (XML Schema Definition) prefixes – point to currently fictitious URIs and specific values of the respective Value Domains' data-format attributes.

The predicates are essentially the concatenated references to the metadata concepts of the ISO/IEC 11179 model described in Section 5. In Fig. 4 for example, the predicate:

```
encr:personWithCancer_TumourPrimarySite_TumourCode
ECISv1
```

is the URI associated with the CDE similar to that described earlier with OC of "Person with cancer" and Property of "primary site of cancer". In this CDE however, the VD "TumourCodeECISv1" would

Indicator	Sex	Cancer	Year	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+
I	M	Lung	2009	0	0	0	0	0	0	6	0	3	34	51	120	148	297	367	504	565	598

Fig. 3. Example of one single row of a CR's aggregated core data set, with column header information.

```
@prefix ecis: <https://ecis.jrc.ec.europa.eu/resource/> .
@prefix ecis-code: <https://ecis.jrc.ec.europa.eu/resource/code/> .
@prefix encr: <http://encr.eu/resource/cde#> .
@prefix pbpr: <http://pbpr.eu/resource/cde#> .
@prefix qb: <http://purl.org/linked-data/cube#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sdmx-code: <http://purl.org/linked-data/sdmx/2009/code#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

ecis:record-1-12 a qb:Observation ;
  encr:personWithCancer_TumourPrimarySite_TumourCodeECISv1 ecis-code:cancer-site-15 ;
  pbpr:caseRegistration_HistoricalYear_YearYYYY "2009"^^xsd:gYear ;
  pbpr:caseRegistration_Indicator_CodeX pbpr:Incidence ;
  pbpr:caseRegistration_NumCases_Number "120"^^xsd:nonNegativeInteger ;
  pbpr:patient_AgeGroupAtDiagnosis_5YrAgeBracketNum pbpr:FiveYrAgeGroupID12 ;
  pbpr:person_SexAtBirth_CodeX sdmx-code:sex-M ;
  qb:dataSet ecis:dataset-1 .
```

Fig. 4. Example of a data entry in a CR's aggregated core data set in RDF, using the Turtle [45] serialisation.

have the list of possible values as defined currently by ECIS [44].

A tentative full RDF definition of the subjects, predicates, and objects used in this PBPR example is provided in [46].

7.1.2. Definition of CDEs

The remaining five predicates in Fig. 4 refer to the other CDEs needed for describing the data associated with the ENCR aggregated core data set, namely:

```
pbpr:caseRegitration_HistoricalYear_YearYYYY
pbpr:caseRegitration_Indicator_CodeX
pbpr:caseRegitration_NumCases_Number
pbpr:patient_AgeGroupAtDiagnosis_5YrAgeBracketNum
pbpr:person_SexAtBirth_CodeX
```

where once again each predicate forms a concatenation of OC, Property, and VD of the CDEs according to the ISO/IEC 11179 model. Since most of the metadata types are generic to all PBPRs, they would

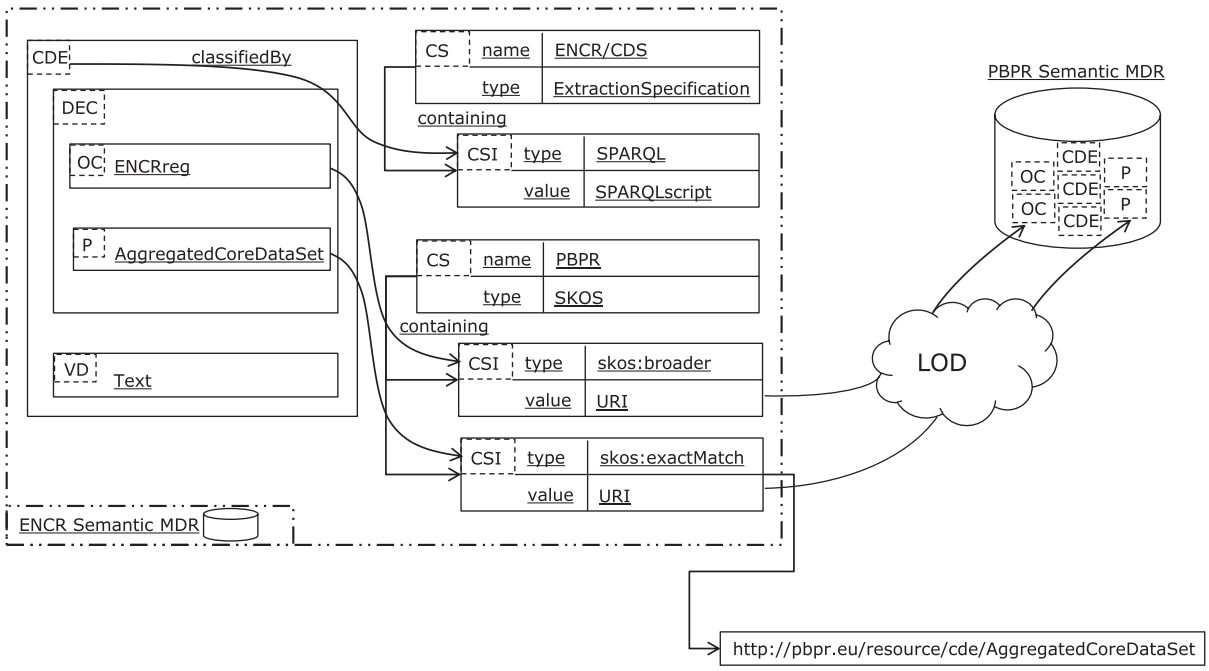

Fig. 5. Semantic links of a CDE and its Property inside the ENCR semantic MDR. The Property (P) is annotated through the SKOS mapping property "exactMatch" to indicate that the CR core data set is an aggregated core data set of a population-based patient registry. The Object Class (OC) is annotated through the SKOS semantic relation "broader" to indicate that a PBPR is related to an ENCR registry but is broader in context. The CDE has an "Extraction Specification", which in this example is a SPARQL script that is defined in Fig. 6.

best be defined and maintained within a generic-PBPR semantic MDR that could then be referenced by different PBPR domains as is illustrated in this example.

7.1.3. Composite CDEs and ISO/IEC 11179

ISO/IEC 11179 defines a data element as a "unit of data that is considered in context to be indivisible", where context is defined as "the circumstance, purpose and perspective under which an object is defined or used" [31]. The standard provides the example of a telephone number that may be considered indivisible in one context but divisible in another (where telephone numbers need to be divided into country code, area code, and local number). To all intents and purposes, the core data set can be considered indivisible for the purpose of making it searchable and accessible as a whole. Not only are the core data sets standardised and harmonised within any given PBPR domain, but an individual field within the aggregated data set is not meaningful without explicit reference to the values of all the other fields in the same row. The way in which the core data set will eventually be used forms a higher-level context outside our immediate concern and since our interest is in allowing access to and retrieval of the core data sets as a whole within each PBPR domain, we specifically define a CDE to represent a registry's aggregated core data set. This marks a slight departure from the aim of [30] where the purpose was to extract the value of specific CDEs associated with the EHRs of specific patients. The alternatives (none of which are preclusive to the above) would be: (a) to specify CDEs representing each individual record within the aggregated core data set, e.g. the number of incident, male (at birth), lung cancer cases aggregated within the age range 55–59 years diagnosed in the year 2016. This however would require a major initial outlay of effort in defining the whole set of CDEs (for each individual tumour type, age bracket, indicator type, sex, and year) as well as complicate the task of reconstituting the entire aggregated core data set for users requiring it; or (b) to provide a user interface for selecting the particular fields of interest within the data set and on the basis of these choices, to run a script on the data set to return the relevant result. This option also would require more initial effort, although such an interface would be useful also in accessing individual record data (discussed in Section 7.3).

7.1.4. Proposed definition for an ENCR aggregated core data set CDE

Following a similar semantic-MDR schematic representation to that provided in [30], Fig. 5 illustrates the definition of a proposed CDE for the ENCR aggregated core data set with semantic links through the LOD cloud.

The CDE is the association of an OC that represents the concept of a population-based ENCR registry with a Property referring to the feature of an aggregated core data set, and a VD that specifies the form of the data element (in this case RDF-formatted text).

The Object Class of the CDE is annotated with a concept of a population-based patient registry (via an association with a classification scheme item) through the SKOS (Simple Knowledge Organization System) [47] semantic relation "broader" to indicate that a PBPR is related to a population-based ENCR registry but broader in context.

The CDE also has an extraction specification specified with a SPARQL script (described in Fig. 6) that can be used to retrieve ENCR-conformant aggregated core data sets from local CR MDRs.

Executing the script in Fig. 6 with the URI of a local registry's aggregated core data set would return all the records on a row-by-row basis with the individual data fields of each record aligned under the column names entitled by the predicates of the RDF triples. The latter are no less than the links to the CDE metadata describing the associated data fields and may be dereferenced to provide all the associated sematic links. Consequently the user has all the information needed to ascertain in a relatively straightforward manner the full semantic meaning of each record.

In a more general scenario, a user might be interested in finding all the available population-based registries having aggregated core data sets. To do this the user would search through the semantic MDR framework on the Property "AggregatedCoreDataSet". The search would retrieve the URIs to all the aggregated core data sets linked to this Property for all PBPRs accessible in the framework. Refining the search on Object Class would then provide the means of classifying the URIs in terms of their specific domains (e.g. CR domain). Finally, the DEC together with the Value Domain define the CDE from which the extraction specification can be found for the specific registry domains. The user then has all the means by which to extract the data from all the local registries within each patient-registry domain.

Even achieving this result would in itself constitute a major milestone in the path towards data interoperability between registries and open up for the first time the means of accessing data from all the various PBPR networks without a major effort on behalf of the user. As an example, it would then be possible to look for potential correlations at population level between indicators relating to pancreatic cancer and type II diabetes in different geospatial regions and ascertain possible cohorts for further high-resolution analyses.

7.2. Use case #2: Access to aggregated full data sets

The core data sets, as useful as they are, hold only a fraction of the data potentially available. Access to the full set of a registry's data

```
PREFIX encr: <http://encr.eu/resource/cde#> .
PREFIX pbpr: <http://pbpr.eu/resource/cde#> .

SELECT *
FROM NAMED <<localRDFfile>>
WHERE {
  ?rowId pbpr:caseRegistration_Indicator_CodeX ?indType ;
    pbpr:person_SexAtBirth_CodeX ?sex ;
    encr:personWithCancer_TumourPrimarySite_ECISv1 ?ca ;
    pbpr:paseRegistration_HistoricalYear_YearYYYY ?year ;
    pbpr:patient_AgeGroupAtDiagnosis_5YrAgeBracketNum ?ageGroup ;
    pbpr:caseRegistration.NumCases.Number ?nos .
}
```

Fig. 6. SPARQL script to retrieve an aggregate core data set from local CRs. The ≪localRDFfile≫ tag is a generic tag that is overwritten by the URI of the RDF graph containing the local data set. The latter is returned after searching the federated semantic MDR framework for links to the "ENCRreg.AggregatedCoreDataSet.Text" CDE.

variables would not only provide users with much richer data sets but also serve, by wider use of the data, to accelerate the data-harmonisation process.

Due to the currently limited degree of harmonisation of the full variable sets for many PBPRs, the CDEs of the non-harmonised variables and the extraction specifications for the full data sets would need to be provided and maintained by the local registries until such time as the variables became harmonised.

Fig. 7 illustrates how the CDE for the full data set could be defined in the local MDR (CaRegXXX): The CDE of Fig. 7 is associated with three CS resources, one used by the CDE itself of type ExtractionSpecification for eventual extraction of the data set; and two of type SKOS - one used by the OC of the CDE, and one used by the Property component of the CDE. The OC is again used to refer to the concept of a population-based ENCR registry through the SKOS mapping property "exactMatch". The CSI value field of the OC's associated CS resource contains the URI of the corresponding OC in the proposed ENCR semantic MDR (http://encr.eu/resource/cde/ENCRreg). The Property refers to the concept of a population-based aggregated full data set also through the SKOS mapping property "exactMatch". The CSI value field of the Property's associated CS resource contains the URI of the corresponding Property in the proposed PBPR semantic MDR (http://pbpr.eu/resource/cde/AggregatedFullDataSet). The VD specifies the form of the data element (which could be either RDF-formatted text or SQL-formatted text), and extraction specification for the aggregated full data set could accordingly be specified either in terms of SPARQL or SQL, depending on the local MDR's decision. To complete the picture, the OC of the CDE in the ENCR semantic MDR (to which the OC in the local CR links) is itself linked to the corresponding OC in the PBPR semantic MDR but this time through the SKOS semantic relation "broader" to show that a PBPR has a broader scope than a CR.

By searching on the Property "AggregatedFullDataSet" over the federated semantic MDR framework, the user would retrieve the URIs to all the full data sets from any associated population-based patient registry. The categorisation of patient-registry domain would again be determined from the OC. The DEC and VD together define the CDE from which the extraction specification can be found. Running the extraction specification defined in the CDE of the local registry would retrieve the aggregated full data set stored in the local MDR.

It is perhaps important to add that whereas the semantic linkages of the CDEs would provide comprehensive descriptions of the CDEs themselves as well as their potential relation to standard terminologies and dictionaries, this may still not be sufficient to provide users with a full understanding of the data paradigm and the interdependency of the data variables themselves. For a higher-level view, it may be necessary to provide a data model and/or ontology describing the data domain. The description of such models can however be integrated into the semantic MDR framework using the mechanism of the Classification Scheme described earlier. As long as the data model/ontology were accessible via URIs, the DEC (constituting the local registry OC and the aggregated full data set Property) for the aggregated full data set could be classified by CSIs that would point to the associated URIs. The higher-level descriptions would then be accessible along with all the other semantic links once the URI of the CDE for the aggregated full data were dereferenced.

7.3. Use cases #3 and #4: Access to record level data and aggregation on demand

Access to all the data variables at record-level would allow the greatest use and value of patient-registry data, but requires explicit patient consent under the EU's recent general data protection

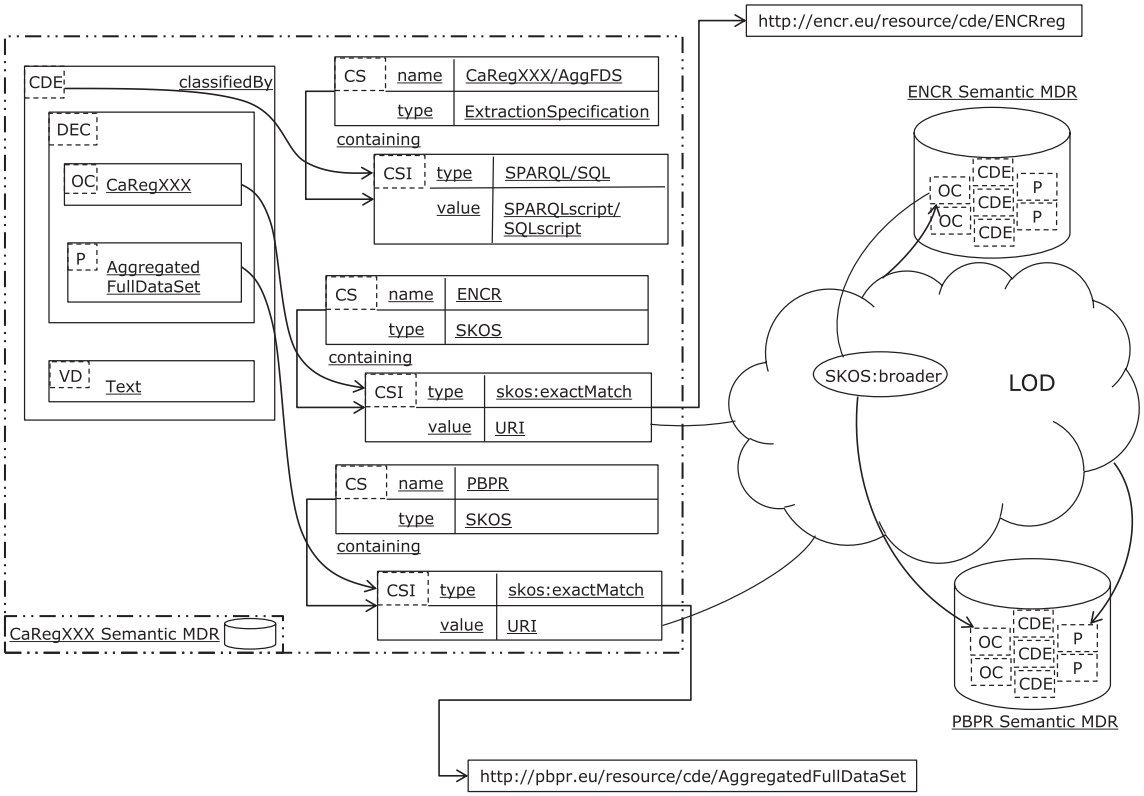

Fig. 7. Semantic links of a CDE inside the local CR (CaRegXXX) semantic MDR. The Object Class (OC) is annotated through the SKOS mapping property "exactMatch" to indicate that the local registry is an ENCR registry—which is itself mapped to the PBPR OC via the SKOS semantic relation "broader" (c.f. Fig. 5). The Property (P) is also annotated through the SKOS mapping property "exactMatch" to indicate that the CR aggregated full data set is an aggregated European harmonized full data set. The CDE has an "Extraction Specification" that, dependent on local decisions, could be a SPARQL or SQL script to return the data set in a way similar to that described for the aggregated core data set.

regulation (GDPR) [48]. The record-level data are considered sensitive data even though they are generally pseudonymised through the re-coding of the patient-identity field.

It should be re-emphasised that the ultimate reason why epidemiological studies require access to individuals' data is for the purpose of selecting the relevant cohorts of patients for testing a particular research hypothesis. Once the cohort is created, the analysis generally proceeds without further reference to individual patients. With this concept in mind, there are potentially two ways in which the data-sensitivity aspect of record-level data might be relaxed and allow data users more straightforward access to the underlying data:

1. *By allowing users to specify the exact criteria for aggregating data*. As an example, a user could ask the registry for the group of patients with survival less than a certain length of time suffering from a given cancer- type. As long as the number of corresponding patients were greater than a pre-defined minimum to prevent possible identification, the returned data set would be an aggregated measure and therefore essentially anonymous. This is essentially the scenario described in Section 4.4.
2. *By providing users pseudonymised individual records with a minimal set of data fields*. Minimising the number of data fields complicates the task of identifying a particular individual. This could be accomplished using a SPARQL/SQL front-end allowing users to specify search criteria based on a number of specific predicate/column names up to a permissible maximum number. Notwithstanding, appropriate measures would have to be in place to avoid successive calls on the same data being able to reconstruct the complete set of variables for any given record.

Data-access procedures for both these scenarios could in principle be automated since the results returned are arguably anonymised data. Requests of this nature could possibly be constructed via a similar type of interface to that described in [32] albeit with extra functionality for handling service negotiation for data retrieval through a firewall.

7.4. Use cases #5: Trace-back to primary data sources

Use case 5 lies somewhat outside the immediate focus of this paper, but it addresses a topic that could bring significant cost-savings to high-resolution studies. Particularly relevant is the work of Sinaci and Laleci Eturkmen [30] in which they developed the semantic MDR framework for EHRs and consequently demonstrated how it is possible to retrieve data for a particular patient held in various clinical systems.

IHE-RFD and FHIR could potentially also provide alternative solutions for this task. Work is in progress to address the cost-benefit of implementing FHIR in the clinical registry world [49]. Whereas FHIR could greatly facilitate the task, it would require widespread uptake and even then may not solve the issue of data stretching back many years held in legacy systems.

The advantage of the semantic MDR framework is that it does not require adherence to any one single standard. It would however require effort to establish the mapping to the various clinical sources for each local registry. Nor could the mappings be easily duplicated in other registries given the wide variety of differences between the various regional and national health infrastructures, processes, and data constructs.

8. Related work

Within Europe, a number of initiatives in context of the EU cross-border healthcare directive [50] have been undertaken to improve interoperability of patient registries.

One of the broader initiatives was the PARENT joint action [28], involving several EU Member States and part-funded by the European Commission's health programme 2008–2013. PARENT aimed both to rationalise and harmonise the development and governance of registries for the purpose of facilitating secondary data usage for research and public health. The action developed a set of methodological guidelines to help overcome commonly encountered challenges in the establishment, operation, and maintenance of registries [3].

PARENT also piloted a "registry of registries" (RoR) – essentially foreseen as a web portal providing reliable and up-to-date information about European patient registries' metadata [51].

Within the Biobanking and Biomolecular Resources European Research In- frastructure (BBMRI-ERIC) [52], the MIABIS (Minimum Information About BIobank data Sharing) community standards [53] were developed for supporting biobank data interoperability. MIABIS provides standardised data elements in XML for describing biobanks and as well as data on associated sample collections. A non-exhaustive list of other past and present domain-specific initiatives include: RD-Connect [54], EU RD Platform [55] (rare-diseases domain); EUBIROD [56] (diabetes domain); EUROCISS [57] (cardiovascular domain); EUROCOURSE [58], ECIS [43] (cancer domain); ECFSPR [59] (cystic-fibrosis domain). With the exception of RD-Connect, the CDEs developed within the domain-specific projects are described in PDF files [21–26] and therefore cannot easily be queried or processed, nor are their semantic contexts provided according to the principles of linked open data. RD-Connect uses the COEUS framework [60] to build and provide the semantic context for rare-diseases registries. Given the sheer number of registries and variation in type of data repositories, COEUS facilitates the process towards semantic-knowledge federation. Whereas COEUS provides some of the functionality of the federated semantic MDR framework and eases the transition of traditional data storage mechanisms towards semantic web technologies, it would be unnecessarily restrictive to force it upon all registry domains. Furthermore, COEUS derives the predicates of the RDF triples from the column names of the underlying data sources and then maps them to the relevant predicates in a given ontology, wherein can lie potential inconsistencies as discussed in [61]. In contrast, the federated semantic MDR framework has these mappings already established within the local MDRs and linked at the different conceptual levels of a data element through the classification scheme associations, allowing greater flexibility in semantic searches. However, some of the shared similarities could potentially be used to provide the semantic mappings to access data elements between the frameworks. Full alignment would be possible by defining the CDEs in terms of ISO/IEC 11179 and then using the semantic MDR framework to link the CDEs to the associated semantic mappings created using COEUS.

9. Discussion

Overcoming the barriers to secondary data usage of PBPRs is an important goal. PBPRs provide a rich source of summarised health data in well defined populations stretching back many years.

Given the complex array of healthcare infrastructures and health data systems and the need still to interface with legacy systems, it is unlikely that any single health data standard will solve all the interoperability issues.

The intention of this work has been to show at a practical level how the federated semantic MDR framework might provide an elegant solution for addressing many aspects of the interoperability challenges facing PBPRs. The framework is able to operate across standards enabling data linkage between disparate systems. It can provide semantic linkage across heterogeneous PBPR domains and is not disruptive. It would also encourage federation of data and thereby remove the need for centralised data collection processes.

Its implementation is however not without certain obstacles.

9.1. Drawbacks of the semantic MDR framework

In order to function across PBPR domains, it would be important to agree some common CDEs at an inter-PBPR domain level. These CDEs

were discussed in Sections 7.1 and 7.2 for concepts relating to aggregated core data sets, aggregated full data sets, and other commonly shared terms. The CDEs would also need to be maintained at this level, which in practice may be difficult to ensure – especially since there is no entity currently operating at this scale of coordination.

Each metadata and data provider would need to establish a local semantic MDR and set up the RESTful services discussed in Section 6.1. For small registries or registries with limited funding, the provision of such services may be a tall order. Moreover, there is a cost in establishing and maintaining the semantic mapping between CDEs. In the most comprehensive scenario, this mapping would have to be performed by the local registry. Leveraging the work already done by existing terminology and classification systems can serve to attenuate the costs and maintenance is perhaps less of an issue given that the data variables recorded by PBPRs do not change that often (records may go back many tens of years and data must remain compatible for time-trend analyses).

Mapping also suffers the drawback of potential loss of information – for example when a data element of broader scope is mapped down to one of narrower scope. The framework avoids such loss of information by using SKOS mapping concepts in which the semantics of these broader or narrower relationships are retained and does not force a one-to-one mapping between data elements where it does not exist. The latter is an important aspect since it furnishes data users with a full picture of the difference between CDEs in different data sets, thereby providing them the necessary information on which to make an informed decision on how to compare/integrate the data. It does however put the onus on the data users to perform the higher level mapping to integrate non-harmonised CDEs from several data sets, which in practice may be difficult to accomplish. It is likely that a dedicated application is needed to help marshal all the data from various sources.

Nevertheless, considering the benefits the framework promises and the fact that all other endeavours toward interoperability do not scale across all registry domains, the steps to implement it are meritorious of the required effort.

9.2. Summary of steps needed to implement a PBPR semantic MDR framework

The steps to implement the framework in the domain of PBPRs can be summarised by the following:

1. Within the PBPR generic domain: to create a number of abstract ISO/IEC 11179 classes (in particular Object classes and Properties) that can be re-utilised at specific registry domain level. Examples would include “Population-Based Patient Registry” as an Object Class, and Properties such as: “Aggregated Core Data Set”, “Aggregated Full Data Set”, “Indicator Type”, “Sex At Birth”, “Year”, “Geolocation Code”, etc. These would provide the hooks for finding all the associated CDEs across different PBPR domains
2. Within the specific patient-registry domain:
   (a) to create the required set of CDE definitions according to ISO/IEC 11179. One example for the CR domain would be the CDE definition for “Cancer Type”, as discussed earlier. Metadata at this level will already exist – at least for describing the variables used within the core data sets – but may not be available in machine-readable ways. In order to reduce the effort, the initial focus could be limited to the definition of the core data variables;
   (b) to set up the semantic links (via the Classification Scheme Item of ISO/IEC 11179) in the manner described by the semantic MDR framework to any relevant standard terminology systems (providing data dictionaries, classification schemata, ontologies, etc.) to which the CDEs are related. Extraction specifications will also need to be provided for the aggregated core and full data sets in the manner described in this paper. The extraction specifications are used to extract the associated aggregated data sets from the local patient registries;
3. Within the local registry/central registry domain: to provide the aggregated core data sets in RDF format;
4. For all MDRs: to establish a RESTful interface enabling the services foreseen by the semantic MDR framework (namely, SPARQL endpoint, CDE endpoint, CDE search, Semantic links, and Extraction specification).

As an initial approach, if the central coordinating entities of the patient- registry domains store the aggregated core data sets as a type of proxy for the local registries, then the central semantic MDR of the different PBPR domains could service all the requests to retrieve the individual core data sets. However, to unlock the full power of the registry data, each individual local registry would also need to set up their own semantic MDR and RESTful interface in order to handle the requests on the local registry. As a result, all data sets would reside on the servers of the local patient registries without the need for any central-level repository; furthermore, automatic access to non-sensitive aggregated full variable-set data would then be possible.

In view of the fact that the implementation of these steps will share many commonalities between registries and also for ease of rolling out such a framework encompassing all population-based registries, it would be worthwhile to prototype the whole concept on one patient-registry domain and, in so doing, create an implementation manual for other domains to follow.

9.3. Cancer registry domain as a suitable prototype

The CR domain would serve as a good starting point. The CR domain is well established and comprises over 200 individual registries. Currently the core data sets are collected centrally and thereafter cleaned and aggregated prior to being made available on the ECIS website [43]. The aim is eventually to eliminate the central data-collection process altogether, thereby avoiding extra overheads and delays as well as the need for retaining copies of data sets with all the consequent maintenance and data-integrity issues. One of the hurdles to overcome before this becomes a reality however concerns the data-validation operation. Data validation is necessary at the central level to ensure all the data sets conform to a similar degree of data quality such that they can more accurately be compared. Work is in progress to provide open-source data-validation software tools based on a standard ENCR data model using an ontological approach, which would allow a federated approach also to the data-validation process. One acid test will be to ascertain if these tools in conjunction with the framework itself provide the necessary robustness to devolve the current prerequisite central processes to the local level. In any case, self-regulation strongly motivates conformance to standard procedures and such a process could well be encouraged via the allocation of data-quality stamps to distinguish between different degrees of quality of data sets. Where data are seen to be essential, more effort will be given to ensuring compliance to standard practices.

10. Results and conclusions

In view of overcoming the issues facing PBPRs regarding secondary usage of data and data-linkage across heterogeneous patient registries, the federated semantic MDR framework provides a powerful, versatile, and – more importantly – non-disruptive solution. The framework maps local metadata to standard metadata descriptions and linking their components semantically via knowledge organisation system ontologies and terminology systems without enforcing compliance to any one common data model. In a world where data have long been collected and managed with local contexts in mind, this is a critical aspect towards allowing secondary data usage without requiring fundamental changes to existing data sets or local data-collection practices.

A major advantage of the framework lies in the fact that it is not a disruptive technology but rather provides the means, via the integration of a number of powerful tools and standards, to link data that would otherwise remain fragmented.

Whereas the implementation is not cost-neutral – a number of elements need to be established and thereafter maintained – and registries already contending with limited resources would undoubtedly require support, the following points must be borne in mind:

- Population-based patient registries contain valuable and important data stretching over many years. The value of the data may be gauged from all the previous initiatives and endeavours to make them interoperable. If a registry is established, it makes inherent sense to ensure the data it collects are interoperable with those of other registries to provide extra value;
- Not agreeing a common framework only postpones the problem of data inter-linkage to some future date – data inter-linkage will always depend on semantic description of the data and where this has not been considered at an early stage, it will have to be done later on and at potentially greater cost;
- Many on-going efforts focused on bringing interoperability across registries purely within a specific-patient domain are not resource-neutral and require considerable effort at all stages of the work. These resources could be re-directed;
- Agreeing a common approach brings economies of scale that can ease the burden on any one registry. Automated tools developed in the course of implementing the framework in one domain can be used to facilitate its implementation in other domains. The initial attention within an individual patient-registry domain need focus only on the critical aspects of metadata definition and semantic linkages;
- An infrastructure based on a framework that is designed for interoperability of EHRs means that interoperability between patient registries and EHRs is factored in from the start. The link to a patient identifier opens up a wealth of possibilities both for clinical processes and clinical research. On the one hand, it would ease the task of patient registries acquiring and validating data and on the other, any high-resolution study addressing a research question motivated from registry data would automatically have the link to many other related health details with which to select the pertinent patient cohort. The current difficulty in gaining access to this information is a cause of considerable effort, expense, and delay in such research;
- The framework can be implemented in degrees without breaking any of the underlying data processes. Metadata are not changed but rather mapped to standard metadata descriptions. Moreover, focusing initially on access to core data sets held centrally by each PBPR domain allows a quick win that can be rolled forward gradually to extend accessibility to other data variables;
- The framework provides a clear model and set of procedures for guiding the establishment of new patient registries and patient registry domains in order to make them compliant from the outset and thereby save future effort in making them interoperable;
- The issues raised within the PARENT guidelines [3] relating to registry data re-use (e.g. data compatibility and comparability; data exchange; mapping of classification codes; and data semantics) are all addressed by the functionality provided by the federated semantic MDR framework. In particular, metadata will be described in a formal manner, removing ambiguities and duplication of terms. Furthermore, it will be described in machine-readable terms and conform to the ISO/IEC 11179 metadata registry standard [31];
- Within the federated semantic MDR framework the need for any overarching registration function (such as PARENT's concept of an RoR) would be redundant – all the metadata are already registered and linked in the framework and therefore can be browsed on a patient-domain basis using a tool similar to the one described in [32];
- The framework would eventually negate the need for any central collections of data for data-validation and data-cleaning needs, thereby avoiding further resource-intensive operations;
- With such a framework in place, attention and resources can be directed to developing user-interface tools for facilitating browsing, searching, marshalling, and fusing data retrieved from multiple data sources. These tools would serve as a front end to what would constitute a research infrastructure in its own right. Machine-readable metadata with their associated semantic descriptions and SKOS linkages to other resources would make the whole registry community amenable to the growing number of AI tools and applications and add the registry data sets to the growing pool of linked open-data resources.

CRediT authorship contribution statement

Nicholas Nicholson: Conceptualization, Methodology, Writing - original draft. **Andrea Perego:** Methodology, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R.E. Gliklich, N.A. Dreyer, M.B. Leavy, (Eds.), Registries for Evaluating Patient Outcomes: A User's Guide [Internet]. 3rd edition. Rockville (MD), Agency for Healthcare Research and Quality (US), Apr. 2014, Chapter 1: Patient Registries. Available from: https://www.ncbi.nlm.nih.gov/books/NBK208625/ (accessed 21 April 2020).
[2] B.C. Drolet, K.B. Johnson, Categorizing the world of registries, J. Biomed. Inform. 41 (6) (2008) 1009–1020, https://doi.org/10.1016/j.jbi.2008.01.009.
[3] M. Zaletel, M. Kralj, Methodological guidelines and recommendations for efficient and rational governance of patient registries, PARENT joint action (cross-border Patient Registries iNiTiative), 2015. https://ec.europa.eu/health/sites/health/files/ehealth/docs/patient_registries_guidelines_en.pdf (accessed 21 April 2020).
[4] J. Olsen, O. Basso, H.T. Sørensen, What is a population-based registry? Scand. J. Public Health 27 (1) (1999) 78, https://doi.org/10.1177/14034948990270010601.
[5] U.S. National Institutes of Health, National Cancer Institute, SEER Training Modules, Types of Registries. https://training.seer.cancer.gov/registration/ (accessed 21 April 2020).
[6] T. Albreht, M. McKee, D.M.Alexe, M. Coleman, J. Martin-Moreno, Making progress against cancer, in: M. Coleman, D.M. Alexe, T. Albreht, M. McKee (Eds.), Responding to the challenge of cancer in Europe. Institute of Public Health of the Republic of Slovenia, Ljubljana, 2008. Chapter 16, pp. 315–327. Available from: https://apps.who.int/iris/handle/10665/107879 (accessed 21 April 2020).
[7] H.T. Sørensen, S. Sabroe, J. Olsen, A framework for evaluation of secondary data sources for epidemiological research, Int. J. Epidemiol. 25 (1996) 435–442.
[8] S. Walters, C. Maringe, J. Butler, J.D. Brierley, B. Rachet, M.P. Coleman, Comparability of stage data in cancer registries in six countries: Lessons from the International Cancer Benchmarking Partnership, Int. J. Cancer 132 (3) (2013) 676–685, https://doi.org/10.1002/ijc.27651.
[9] L. Teppo, E. Pukkala, M. Lehtonen, Data quality and quality control of a population-based cancer registry. Experience in Finland, Acta Oncol. 33 (4) (1994) 365–369.
[10] P. Aspden, J.M. Corrigan, J. Wolcott, et al. (Eds.), Ch.4 Health Care Data Standards, in Patient Safety: Achieving a New Standard for Care, National Academies Press (US), Washington (DC), 2004, pp. 127–168.
[11] Integrating the Healthcare Enterprise (IHE), IHE Profiles. https://www.ihe.net/resources/profiles/ (accessed 15 April 2020).
[12] HL7 International, C-CDA (HL7 CDA R2 Implementation Guide: Consolidated CDA Templates for Clinical Notes - US Realm). http://www.hl7.org/implement/standards/product_brief.cfm?product_id=492 (accessed 21 April 2020).
[13] Health Level Seven (HL7), The Relationship between FHIR and other HL7 Standards. https://www.hl7.org/fhir/comparison.html (accessed 21 April 2020).
[14] Health Level Seven (HL7), Introducing HL7 FHIR Release 4. https://www.hl7.org/fhir/summary.html (accessed 21 April 2020).
[15] Integrating the Healthcare Enterprise (IHE), IHE Cross-Enterprise Document Sharing. https://wiki.ihe.net/index.php/Cross-Enterprise_Document_Sharing (accessed 21 April 2020).
[16] World Health Organization (WHO), Classification of diseases ICD. https://www.who.int/classifications/icd/en/ (accessed 21 April 2020).
[17] A. Awaysheh, J. Wilcke, F. Elvinger, L. Rees, W. Fan, K. Zimmerman, A review of medical terminology standards and structured reporting, J. Vet. Diagn. Invest. 30 (1) (2018) 17–25, https://doi.org/10.1177/1040638717738276.

[18] National Institutes of Health (NIH), National Library of Medicine, Unified Medical Language System (UMLS). https://www.nlm.nih.gov/research/umls/index.html (accessed 21 April 2020).
[19] Observational Health Data Sciences and Informatics (OHDSI) forum, International Classification of Diseases for Oncology (ICD-O). https://forums.ohdsi.org/t/international-classification-of-diseases-for-oncology-icd-o/1851 (accessed 23 April 2020).
[20] M. Valentic, B. Plese, I. Pristas, D. Ivankovic, Addressing the data linking challenges: interviewing for best practices in patient registry interoperability, Methods Inf. Med. 56 (2017) 407–413, https://doi.org/10.3414/ME16-02-0029.
[21] European Cystic Fibrosis Society (ECFS), Common data elements metadata, cystic fibrosis. https://www.ecfs.eu/projects/ecfs-patient-registry/Variables-Definitions, 2019 (accessed 13 April 2020).
[22] European Cardiovascular Indicators Surveillance Set (EUROCISS), Common data elements metadata, cardiovascular. http://www.cuore.iss.it/eurociss/en/indicators-eu/indicators_europe.asp, 2019 (accessed 13 April 2020).
[23] European Commission, Common data elements metadata, congenital anomalies. https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/2.2.1b_28_Dec2018.pdf, 2019 (accessed 9 September 2019).
[24] European best information through regional outcomes in diabetes (EUROBIROD), BIRO data elements. http://www.eubirod.eu/eubirod_DataStandards.htm (accessed 21 April 2020).
[25] European Commission, Common data elements metadata, rare diseases. https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements, 2019 (accessed 9 September 2019).
[26] C. Martos, E. Crocetti, O. Visser, B. Rous, F. Giusti, et al, A proposal on cancer data quality checks: one common procedure for European cancer registries (version 1.1), JRC Technical Report. Publications Office of the European Union, 2018. https://doi.org/10.2760/429053.
[27] Union for International Cancer Control's (UICC), TNM classification of malignant tumours. https://www.uicc.org/resources/tnm (accessed 21 April 2020).
[28] European Commission Consumers, Health, Agriculture, and Food Executive Agency (CHAFEA) Health Programmes Database, Cross-Border Patient Registries Initiative [PARENT] [20112302] - Joint Actions. Project Summary, https://webgate.ec.europa.eu/chafea_pdb/health/projects/20112302/summary/, 2015 (accessed 9 September 2019).
[29] Orphanet, Orphanet Report Series, Rare Disease Registries in Europe. http://www.orpha.net/orphacom/cahiers/docs/GB/Registries.pdf, 2019 (accessed 21 April 2020).
[30] A.A. Sinaci, G.B. Laleci Erturkmen, A federated semantic meta- data registry framework for enabling interoperability across clinical research and care domains, J. Biomed. Inform. 46 (2013) 784–794, https://doi.org/10.1016/j.jbi.2013.05.009.
[31] ISO/IEC, 2019. ISO/IEC 11179: Information technology – Metadata Registries (MDR) Parts 1-7, ISO Standards. International Organization for Standardization. http://metadata-standards.org/11179/, 2015 (accessed 21 April 2020).
[32] A.A. Sinaci, G.B. Laleci, Erturkmen, Gonul, et al, Postmarketing safety study tool: A web based, dynamic, and interoperable system for postmarketing drug surveillance studies, Biomed Res. Int. (2015), https://doi.org/10.1155/2015/976272.
[33] Integrating the Healthcare Enterprise (IHE), IHE Data Element Exchange (DEX) profile. https://wiki.ihe.net/index.php/Data_Element_Exchange (accessed 15 April 2020).
[34] Australian Institute of Health and Welfare, METeOR: Metadata online registry. http://meteor.aihw.gov.au/, 2019 (accessed 9 September 2019).
[35] World Health Organization, International statistical classification of diseases and related health problems – 10th revision (ICD-10 Version: 2016). https://icd.who.int/browse10/2016/en, 2016 (accessed 21 April 2010).
[36] A. Fritz, C. Percy, A. Jack, et al, International classification of diseases for oncology, 3rd ed. Technical Report, World Health Organization. https://apps.who.int/iris/handle/10665/42344, 2000 (accessed 9 September 2019).
[37] W3C Web Ontology Language (OWL). https://www.w3.org/TR/owl-features/, 2004 (accessed 9 September 2019).
[38] G. Klyne, J.J. Carroll, B. McBride, RDF 1.1 Concepts and Abstract Syntax, W3C Recommendation, World Wide Web Consortium. https://www.w3.org/TR/rdf11-concepts/, 2014 (accessed 9 September 2019).
[39] W3C, Web Services Architecture. https://www.w3.org/TR/ws-arch/ (accessed 24 April 2020).
[40] W3C SPARQL Working Group, 2013. SPARQL 1.1 Overview. W3C Recommendation. World Wide Web Consortium. https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/, 2013 (accessed 21 April 2020).
[41] National Center for Biomedical Ontology, BioPortal. http://bioportal.bioontology.org (accessed 21 April 2010).
[42] M.D. Wilkinson, M. Dumontier, B. Mons, et al., The FAIR guiding principles for scientific data management and stewardship, Sci. Data 3 (2016), https://doi.org/10.1038/sdata.2016.18.
[43] European Commission, European cancer information system (ECIS). https://ecis.jrc.ec.europa.eu/, 2019 (accessed 9 September 2019).
[44] European Commission, List of cancer sites for estimates 2018 (ECIS). https://ecis.jrc.ec.europa.eu/pdf/Estimates_cancer_sites.pdf, 2018 (accessed 9 September 2019).
[45] E. Prud'hommeaux, G. Carothers, RDF 1.1 Turtle – Terse RDF Triple Language. W3C Recommendation. World Wide Web Consortium. http://www.w3.org/TR/2014/REC-turtle-20140225/, 2014 (accessed 21 April 2020).
[46] A. Perego, N. Nicholson, RDF representation of the ECIS data set, v1, OSF, 2020 https://osf.io/ya6bc/.
[47] SKOS Simple Knowledge Organization System Reference, W3C Recommendation. https://www.w3.org/TR/2009/REC-skos-reference-20090818/, 2009 (accessed 21 April 2020).
[48] European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) OJ L 119, 1–88. http://data.europa.eu/eli/reg/2016/679/oj, 2016 (accessed 21 April 2020).
[49] S. Blumenthal, Improving Interoperability between Registries and EHRs, AMIA Jt Summits Transl. Sci. Proc. 18 (2017) 20–25.
[50] European Union, Directive 2011/24/EU of the European Parliament and of the Council of 9 March 2011 on the application of patients' rights in cross-border healthcare. OJ L 88, 45–65. http://data.europa.eu/eli/dir/2011/24/oj, 2011 (accessed 21 April 2020).
[51] V. Pajić, T. Čebular, M. Kostešić, Pilot Registry of Registries (RoR) Phase 1 Development Report. Technical Report. PARENT project. https://webgate.ec.europa.eu/chafea_pdb/assets/files/pdb/20112302/20112302_d06-00_en_ps_pilot_registry_of_registries_phase_1_development_report.pdf, 2103 (accessed 21 April 2020).
[52] BBMRI-ERIC, Biobanking and biomolecular resources European research infrastructure (BBMRI-ERIC). http://www.bbmri-eric.eu/, 2019 (accessed 9 September 2019).
[53] MIABIS Poject, Minimum information about biobank data sharing (MIABIS) standards. https://github.com/MIABIS/miabis/wiki, 2019 (accessed 9 September 2019).
[54] CORDIS, RD-CONNECT: An integrated platform connecting registries, biobanks and clinical bioinformatics for rare disease research. Project Fact Sheet. European Commission. https://cordis.europa.eu/project/id/305444, 2019 (accessed 9 September 2019).
[55] European Commission, European platform on rare disease registration (EU RD Platform). https://eu-rd-platform.jrc.ec.europa.eu/, 2019 (accessed 9 September 2019).
[56] EUBIROD Network, European best information through regional outcomes in diabetes (EUBIROD). http://www.eubirod.eu/, 2019 (accessed 21 April 2020).
[57] EUROCISS, European cardiovascular indicators surveillance set (EUROCISS). http://www.cuore.iss.it/eurociss/en/project/project.asp, 2019 (accessed 9 September 2019).
[58] CORDIS, EUROCOURSE – Europe against Cancer: Optimisation of the Use of Registries for Scientific Excellence in re- search. Project Fact Sheet, European Commission. https://cordis.europa.eu/project/id/219453, 2019 (accessed 9 September 2019).
[59] European Cystic Fibrosis Society (ECFS), European Cystic Fibrosis Society patient registry (ECFSPR). https://www.ecfs.eu/projects/ecfs-patient-registry/project, 2019 (accessed 9 September 2019).
[60] P. Lopes, J. Oliveira, COEUS: "semantic web in a box" for biomedical applications, J. Biomed. Semantics 3 (2012), https://doi.org/10.1186/2041-1480-3-11.
[61] P. Sernadela, L. González-Castro, C. Carta, et al., Linked registries: Connecting rare diseases patient registries through a semantic web layer, Biomed. Res. Int. (2017), https://doi.org/10.1155/2017/8327980.