We are IntechOpen,
the world's leading publisher of Open Access books
Built by scientists, for scientists

5,800
Open access books available

142,000
International authors and editors

180M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

BOOK CITATION INDEX
CLARIVATE ANALYTICS
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Dotting the "i" of Interoperability in FAIR Cancer-Registry Data Sets

Nicholas Nicholson, Francesco Giusti, Luciana Neamtiu,
Giorgia Randi, Tadeusz Dyba, Manola Bettio,
Raquel Negrao Carvalho, Nadya Dimitrova,
Manuela Flego and Carmen Martos

Abstract

To conform to FAIR principles, data should be findable, accessible, interoperable, and reusable. Whereas tools exist for making data findable and accessible, interoperability is not straightforward and can limit data reusability. Most interoperability-based solutions address semantic description and metadata linkage, but these alone are not sufficient for the requirements of inter-comparison of population-based cancer data, where strict adherence to data-rules is of paramount importance. Ontologies, and more importantly their formalism in description logics, can play a key role in the automation of data-harmonization processes predominantly via the formalization of the data validation rules within the data-domain model. This in turn leads to a potential quality metric allowing users or agents to determine the limitations in the interpretation and comparability of the data. An approach is described for cancer-registry data with practical examples of how the validation rules can be modeled with description logic. Conformance of data to the rules can be quantified to provide metrics for several quality dimensions. Integrating these with metrics derived for other quality dimensions using tools such as data-shape languages and data-completion tests builds up a data-quality context to serve as an additional component in the FAIR digital object to support interoperability in the wider sense.

Keywords: cancer registries, data interoperability, ontologies, description logics data harmonization, data validation, data quality, FAIR data

1. Introduction

Comparison of cancer indicators across different regions and countries is important to understand the effectiveness of cancer prevention and control measures. Considerable care has to be taken however to ensure that the data are indeed comparable and have the necessary level of quality not to result in the production of biased or misleading statistics. Centralized processes to ensure comparability of data are costly in terms of time and resources and should ideally be supported with efficient and effective automated tools. The goal towards the eventual federation of such processes requires the means of formally ascertaining the level of the quality of the underlying data.

1.1 Population-based cancer registries

Population-based cancer registries (CRs) are information systems designed for the collection, storage, and management of data on cancer patients. They collate information on all cancer cases occurring in a defined population and play a critical role in the planning and evaluation of cancer control activities at population level (particularly via trends in incidence, mortality, prevalence, and survival), as well as in identifying good practices of patient care [1, 2]. They also provide the means for evaluating the effectiveness of screening programs and contribute actively to cancer epidemiological research.

CRs may be nationally based, covering the entire country (such as in Europe for Finland, Sweden, and Slovenia), or regionally based (such as in France, Italy, and Spain). Whereas regional CRs may provide total coverage of the country, in some cases they only provide partial coverage and estimations based on the partial coverage are used to provide national statistics. The production of reliable statistics is directly dependent on the quality of the underlying CR data.

1.2 CR data collection and cleaning process

The data collected by a CR are in accordance with the purpose for which the registry has been established, dependent on the available information and resources. Nevertheless, the accent is on the quality of the data rather than on the quantity [3]. Whereas the initial focus was on monitoring cancer incidence and the trends over time, many registries now collect patient follow-up details in order to compute survival.

CRs need to register all cancers diagnosed in a defined area and have consequently to access multiple data sources, including hospital discharge and outpatient records, pathology laboratory results, oncology/radiotherapy/clinical hematology records and death certificates. The combination of such sources is the cornerstone of the data collection process [4]. Additional data sources include screening programs, communications from general practitioners, drug prescriptions, and insurance reimbursement claims.

Sets of rules and linkage routines are normally used to create provisional incidence records, which are then verified within a few months to confirm or discard cases [5]. Once the incidence data set has been consolidated, the data are thereafter cleaned according to specific data-cleaning rules. Additional to the local CR procedures, wider standards for data collection, coding, reporting, and validation are required to facilitate data interoperability. Such standards are generally defined and agreed at national or transnational level, especially in relation to the data comprising the base denominator or the common data set.

1.3 Importance of CR data harmonization

Within the last couple of decades, CR data have improved dramatically in quality and quantity, due largely to technological advances and the improved means for reliable record linkage [6, 7]. Owing to the fact that CRs collect and integrate data from very heterogeneous multiple information sources, a process of data harmonization is required both preceding and following linkage according to national and internationally accepted procedures. This process of harmonization has been defined as "all efforts to combine data from different sources and provide users with a comparable view of data from different studies" [8] and is a critical element for accurate and meaningful inter-comparison of CR data. It is also extremely

important for the correct usage of anonymized or aggregated CR data in secondary-data analyses [9].

An example of the importance of CR data harmonization relates to the implementation of the 1995 European Network of Cancer Registries' (ENCR) recommendations for the coding of bladder tumors in the Scottish CRs in the year 2000. After the introduction of the recommendations, bladder tumor incidence rates halved [10] and became similar to those of other registries following the same rules. Notwithstanding such changes in coding, it always remains possible to calculate rates with the previous rules in order to assess time trends.

1.4 CR associations and networks

In the US, the North American Association of Central Cancer Registries (NAACR) develops and promotes uniform data standards for cancer registration. These standardization efforts are of direct importance to the North American Surveillance, Epidemiology, and End Results (SEER) program [11] involving twenty-one North American CRs covering more than one third of the U.S. population.

Within Europe, the standardization efforts of the ENCR, comprising over 150 individual registries, are similarly of importance to the European Cancer Information System (ECIS) [12]. The International Association of Cancer Registries (IACR), the International Agency for Research on Cancer (IARC), the European Commission, and ENCR have all played an essential role in European CR harmonization.

The harmonization efforts ultimately benefit endeavors to compare cancer statistics at the global level [13, 14]. Data harmonization for inter-comparison purposes is generally achieved via the specification of common data sets in which the ranges and interdependencies of a core set of variables are defined by an agreed set of specific rules. The harmonization process is time consuming and requires consultation and agreement across a wide range of stakeholders, especially when the common data set serves multiple purposes. An example of a common data set comprising some fifty data variables and the rules specifying the variable values/ranges and the inter-variable relationships is provided in [15]. The ENCR common data set includes variables related to the patient, the tumor (including stage), treatment, and follow-up.

Owing to the need to ensure a high and consistent level of quality and harmonization, the CR common data sets are currently collected and processed centrally. Whereas centralized processes help control and ensure consistency, they add extra time delays in making the data available – not least from the overheads occasioned by increasingly stricter data-protection paradigms. Data cleaning and harmonization for CR inter-comparison purposes could be made more efficient by devolving the centralized processes to the local level – so long as consistency and data quality can be assured. Conformance of CR data to the FAIR data principles is key to realizing this aim.

1.5 FAIR data principles

The four principles of FAIR data, encompassed in their felicitously named acronym, underlie the need for data to be: findable, accessible, interoperable, and reusable [16], also at a machine-readable and inferable level. The meaning of each term is elaborated by a set of three or four qualifying elements. The challenges to making data FAIR, in terms of the questions that have to be addressed, and some of the mechanisms towards meeting those challenges are summarized in **Table 1**.

FAIR data principle	Questions to address	Possible means for addressing the needs
Findable	Do the data exist and where exactly?	Data catalogs and inter-linkage of catalogs, with relevant search functions; registration of the data under unique identifiers; persistent links and identifiers; searchable metadata; appropriate synonym lists for search terms
Accessible	Is authorization needed to access the data? How can the data be accessed physically?	Data access and user identification controls; authorization request interfaces; application programing interfaces; data extraction scripts; file format metadata; identification of relevant application tools
Interoperable	Can the data be integrated/combined fully/partially with another data set? Can the data be loaded from different applications? Are the data properly comparable with other data? What is the context of the data? How do the variables inter-relate? What are the measurement units of the variables? How can the measurement units be mapped to similar terms in another data set measured in different units?	Metadata descriptions of data variables; linkage of metadata terms to standard data dictionaries; mapping systems; knowledge organization systems; data quality contexts
Reusable	Does the data set contain limitations/disclaimers/assumptions? Are there data restrictions/licenses? Can the data be used for other purposes? Will the data still be accessible at a future date? May the data change over time?	Contextual and provenance metadata; data-usage licenses; data persistence mechanisms; data-maintenance policies

Table 1.
Challenges involved in making data FAIR, some of the questions that have to be addressed, and possible mechanisms for addressing them.

The foundations of FAIR were in fact laid down in several earlier initiatives [17] and the EU is actively supporting activities to progress the underlying concepts. Interoperability is arguably the most challenging of the four FAIR data principles outside of access to personalized data and is discussed further in Section 2. In relation to findable data, health data providers in many countries have started to create data portals and data catalogs.

Whereas a number of international CR portals provide access to anonymized and aggregated CR data sets [11, 12], it is not usually possible to provide secure access to record-level data through automated protocols due to the sensitive nature of health data, although SEER does provide an example of a way to access cancer data following a set of specific conditions. The challenges to CR data accessibility as far as record-level data are concerned are in fact less technical than administrative in view of the legal aspects of data-protection laws. Indeed, they are generic to all data where identification of a person is possible and, even with anonymized data sets, care has to be taken to ensure that persons cannot be re-identified using other data sources. Steps are being taken in the EU, where the data-protection laws are amongst the strictest in the world, to address mechanisms to facilitate authorized access to health data.

Reusability for CR data mainly refers to their use for secondary-data purposes and hinges on accurate and comprehensive description of the data in both the contextual and semantic sense. In this regard, there is a close relationship with the

principle of semantic interoperability (c.f. Section 2.1) – if the data are comprehensively described, the possibility for data reuse is greatly assisted. The latter may be appreciated to some extent by considering SEER data, which are well described in terms of metadata and draw from data adhering to the NACCR data standards. SEER data have consequently led to hundreds of scientific publications on cancer epidemiology. In contrast, the health data environment in Europe is extremely fragmented, but recent initiatives on data reuse are described in [18], including national initiatives in Finland, France, Portugal, and Italy. Within the EU as a whole, the first preparatory steps have been undertaken to create a European Health Data Space (EHDS) [19] for facilitating primary and secondary reuse of health data.

2. Data interoperability

The three qualifying elements defined under FAIR's interoperability principle [16] are in relation to knowledge representation – with particular reference to the use of formal, shared languages and vocabularies as well as linkage to other data descriptors/metadata. Such aspects largely refer to syntactic and semantic interoperability.

2.1 Semantic interoperability

Mechanisms to address semantic interoperability include metadata schemas drawing on standard data dictionaries and thesauri, metadata catalogs (e.g. Data Catalog Vocabulary, DCAT [20]), metadata registries (e.g., ISO/IEC 11179 metadata registry standard [21]), knowledge organization systems (e.g. Simple Knowledge Organization System – SKOS [22]), linked open data (LOD) or any combination of these. Such mechanisms can be incorporated into frameworks and architectures designed for the purposes of supporting FAIR data processes.

A non-exclusive list of FAIR-supporting infrastructures include: beacons [23, 24], used primarily for discovering and sharing of genomic data; a federated semantic metadata registry framework [25], which also provides a potential model for population-based patient registries including CRs [26]; the MOLGENIS data platform for data sharing [27]; the Apache Atlas data governance and metadata framework [28]; the European Open Science Cloud (EOSC) interoperability framework [29]; and the FAIR digital object framework [30]. The way in which the FAIR digital object concept is able to support data interoperability, particularly with reference to EOSC, has been discussed in [31].

The main challenges to semantic interoperability lie in the interlinkage, mapping, and maintenance of metadata between different standards and systems. The availability of standard dictionaries and ontologies together with knowledge organization systems such as SKOS allow data providers to describe their record-level metadata variables in ways meaningful for data users to combine data sets from different data sources. The fact that these standard resources are available in machine-readable ways opens up the possibility for automation of the data-linkage process by intelligent agents, especially when used in conjunction with data registration and cataloging systems.

As important as the semantic context of data is, it does not fulfill all the requirements to make data interoperable. According to the Data Interoperability Standards Consortium [32], data interoperability concerns "the ability of systems and services that create, exchange and consume data to have clear, shared expectations for the contents, context and meaning of that data."

Whereas semantic definitions and linkages of metadata can help describe the context and meaning of data, they cannot per se vouch for the quality of the data. Data quality is of prime importance for CRs whose data are compared between regions and countries for epidemiological purposes or for gauging the effectiveness of cancer healthcare policy initiatives.

2.2 Data quality

Without having some information regarding the quality and veracity of the data sets to be combined, any assumptions drawn from the data integration will at best be speculative. The FAIR data principles do not explicitly address such aspects, apart from in the sense that the usefulness of the data is somehow determinable by the user [33]. One of the qualifying elements under the reusable principle however does require that (meta)data meet domain-relevant community standards, of which quality could arguably form a part, and acknowledgement is given to the critical importance of the quality dimension as identified in the initiatives on which FAIR builds [17].

Various ways for defining data quality have been propounded, particularly in relation to terms of classification/categorization. The ideas build on research conducted in the 1990s, mainly in relation to total data quality management (TDQM) for business processes. An overview of this early work [34] further developed the ideas and formulated a hierarchical data-quality framework in order to addresses the contemporary needs of big data with a view to developing data-quality evaluation algorithms. The hierarchy consists of fourteen elements (with a number of associated indicators) classified under the five dimensions of: availability, usability, reliability, relevance, and presentation quality. Most of these dimensions turn out to be closely aligned with the FAIR data principles and are therefore inherent to the objectives of the FAIR digital object framework (FDOF) [30]. The FDOF provides the means of resolving the identifier associated with a FAIR digital object into sets of information relating to the features required by the FAIR data principles. Factoring out these commonalities essentially removes all but the "reliability" dimension (equating to the trustworthiness of data) in the hierarchy of [34] and one of the elements (Timeliness) under the "availability" dimension as summarized in **Table 2**.

Despite the lack of a universally agreed data-quality system, five of the resulting six elements are common to five of the six quality dimensions identified in [35], which also provides suggested metrics. The different sixth elements are "auditability" and "uniqueness" respectively. In total, the seven quality elements (which we refer to as quality dimensions in line with the terminology used in [35]) are described in **Table 3** together with the proposed means of measurement:

ISO 8000 is an international standard for managing, measuring, and improving the quality of data. Part 8 of the standard [36] (Information and data quality: Concepts and Measuring) can be used independently of the other parts and is specifically focused on providing the means for measuring the quality of data and information against scales that the standard requires the enterprise to establish. It can therefore be used as a means for auditing the data quality.

ISO 8000-8 categorizes data/information quality under: syntactic quality, semantic quality, and pragmatic quality. Syntactic quality relates to the degree in which the data/information conforms to its metadata specifications and the standard requires the specification of a full set of syntactic quality rules. Semantic quality relates to the correspondence/relationships of data or information to other entities as represented in a conceptual model. The standard requires a documented conceptual model and a description of the means used for verification against the

Big data quality dimension	Big data quality element	FAIR principle
Availability	Accessibility	A
	Timeliness	—
	Authorization	A
Usability	Definition/documentation	I,R
	Credibility	R
	MetaData	F,I
Reliability	Accuracy	—
	Integrity	—
	Consistency	—
	Completeness	—
	Auditability	—
Relevance	Fitness	R
Presentation quality	Readability	A,I
	Structure	A,I

Table 2.
Cross-matrix of the quality dimensions (and associated elements) proposed for big-data quality [34] with the different FAIR principles.

Dimension	Measure	Unit of measure
Completeness	Degree in which all the essential data are provided. Can be measured at both data level (missing data records) and variable level (missing variables within a record)	Percentage/ratio (e.g. proportion of captured data against potential of 100%)
Integrity/Validity	Degree in which data types are standardized or conform to rules and relations encapsulated in the data.	Percentage/ratio (e.g. number of non-conformant data elements missing as a ratio of number of records).
Consistency	Differences found for data entities (or their representations) that should be identical or equivalent	Number (e.g. number of differences)
Accuracy	Degree in which the real-life situation is different from its representation	Percentage (e.g. percentage of records to that pass pre-specified data-accuracy rules;
Timeliness	Degree in which the data are representative of the current situation	Time difference
Uniqueness	Redundancy of data which could otherwise be derived, leading to maintenance and consistency issues	Percentage to total of duplicates data/data variables
Auditability	Ease in which/extent to which auditors can evaluate the quality of the data	An agreed or standardized scale

Table 3.
Description and proposed units of measurement of the seven generally agreed data-quality dimensions.

model. Pragmatic quality concerns usage-based requirements that have to be expressed as specific perspectives or dimensions not covered by the other two quality criteria. It can relate to such aspects as accessibility, completeness, security, etc. Using a standard such as ISO 8000-8 would address the issue of auditability as

well as allow the means for formally specifying the other six quality dimensions and the metrics for their measurement.

2.2.1 Quality metrics of CR common data set

Regarding the CR common data set, the metrics related to variable-completeness (i.e. completeness of the common data mandatory variable set), timeliness, and uniqueness can be relatively easily defined. The common data set specifies the permitted set of variables and qualifies which variables are mandatory. Timeliness can be ascertained from the most recent batch of case registration dates, and uniqueness can be addressed by ensuring that the common data-set template does not lead to duplication of data contained in another variable. The more intricate quality dimensions regard integrity, accuracy, consistency, and data-completeness (completeness of the cancer cases within the catchment area of the population).

Whereas integrity and consistency can be assessed from the data, accuracy and data-completeness have to be ascertained from the real-life situation [35]. It is a process followed by CRs when cross checking summary values against data from the primary data feeds (e.g. hospital/clinical records). There may also be accuracy issues within the primary records themselves, such as incorrect data entry, which may be difficult to ascertain at the CR level. Integrity and consistency checks may be able to serve as a proxy in some instances where data entry is incorrect and in violation of the data rules; more subtle, systematic errors could possibly be detected using variances in frequency measures on variables. Establishing a formal data-quality process such as ISO 8000-8 at the first point of data capture is however perhaps the only way in which to assess the steps taken to ensure data accuracy. Such a process if harmonized across the data sources could provide a standard metric to integrate into the quality stamp of further processing operations. Metrics for estimating data completeness of CR data have been summarized in [37]. The data-quality dimensions most relevant to each stage of the CR data throughput chain are depicted in **Figure 1**.

The decision processes underlying the choices to combine data sets dependent on their quality metrics will depend largely on the intended purpose of the end application. The means for one possible decision-making framework is proposed in [38]. The framework is presented in terms of business-related data but raises a number of important considerations. It lays down five requirements for data-quality metrics and argues these requirements in practical examples of metrics proposed by others for measuring the specific quality dimensions of timeliness, completeness, reliability, correctness, and consistency (where correctness corresponds to accuracy and the metric for consistency can be applied also to integrity). The five requirements are:

1. provision of minimum and maximum values;

2. provision of interval-scaled values;

3. means of determining the metric values on the basis of the associated configuration parameters and also whether the quality-criteria objectivity, reliability, and validity of the metric are fulfilled;

4. consistent aggregation of metric values on different data-view levels; and

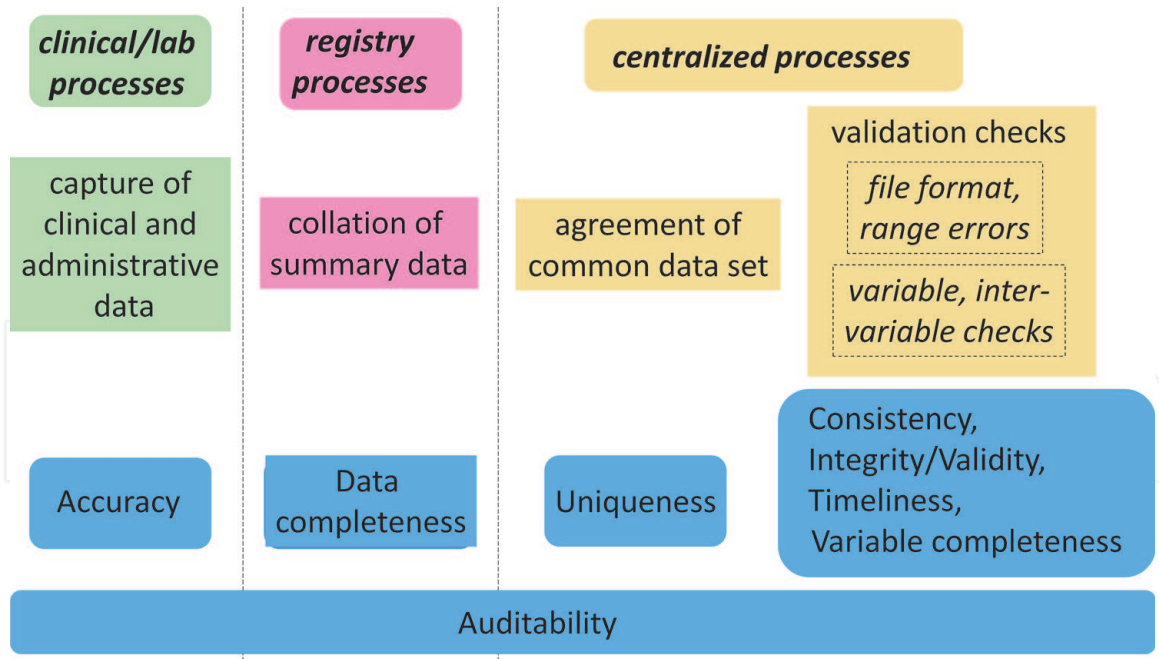5. economic efficiency of the metric (i.e. the cost incurred by the metric).

Figure 1.
Data-quality dimensions relevant to the different stages in the CR common data set throughput process. Auditability can span all processes.

3. Ontologies and underlying foundations on description logics

Ontologies are relevant for describing the semantic relationships between entities in a data model. Bioportal [39] provides a comprehensive repository of biomedical ontologies. The Web Ontology Language (OWL) [40] underlies many of these ontologies and represents the concept definitions and relations between them as sets of Resource Description Framework (RDF) [41] graphs.

Interestingly however, ontologies formulated on description logic (such as OWL) can also be made to provide a basis for ascertaining the quality of data sets. A single tool can thereby be developed to handle both the semantic and the data-quality contexts. Whereas we present a model for achieving this for CR data, the concept is sufficiently generic to be applied to other data domains. An important requirement is that some form of data-validation rules are specified a priori.

For the purposes of comparing CR data, a common data set specifies the metadata of a minimum set variables to be included. Whereas, the availability of a common data set is not necessarily an essential aspect of the data-quality model, it does however aid the process to provide data-quality metrics easily interpretable by the end application.

3.1 Description logics

Description logics (DLs) are a family of languages used to represent in a structured and formal sense knowledge about a given domain [42]. They also provide the means for a degree of machine-reasoning allowing automated inferences to be made on the basis of statements concerning that knowledge.

DL languages are classified by language expressivity. Expressivity basically determines the richness of the modeling capacity of the language; a language with greater expressivity is able to model more complex relationships but at a cost of computing performance. In view of the latter, it is generally preferable to limit the DL expressivity to the minimum needed for the modeled aspects of the domain.

Knowledge about a domain can be captured in an OWL ontology using DL statements that are be classified into TBox and ABox axioms. TBox axioms refer to the terminological part of the ontology and ABox axioms, to the assertional part. The terminological part is analogous to the database concept of a database schema, which describes the structure or layout of the database while the assertional part is analogous to a particular instance or population of a database described by that schema [43]. Thus, OWL TBox axioms describe the hierarchies and relationships between OWL classes and ABox axioms describe specific instances of classes, also referred to as individuals.

The primary two semantic constructs DLs use are: unitary predicates (or concepts) describing entities equating to OWL classes/individuals; and binary predicates (or roles, equating to OWL properties) that describe relationships between entities. DLs are termed as decidable fragments of first-order logic [42] and TBox and ABox statements can in fact be expressed as first-order logic statements. The expressivity of a DL language determines the set of operators permitted. The Attributive Language with Complement (ALC) expressivity allows quite a rich modeling language to handle most of the validation checks in the ENCR common data set. ALC includes: subclasses ($\sqsubseteq$), intersections ($\sqcap$), unions ($\sqcup$), negation ($\neg$), existential restrictions ($\exists$), and universal restrictions ($\forall$). The restriction operators are used for qualifying the entities on which a given role acts, with $\exists$ specifying the notion of an "at-least-one relationship" and $\forall$ the notion of an "only relationship" and are similar to the existential and universal quantifiers of first-order logic.

3.2 Transcribing the data model and validation rules in DL

The data-validation rules encapsulate the part of the domain model that minimally needs to be modeled. The challenge lies in designing the ontology in a way that is straightforward to understand, easy to maintain, and models the data relationships satisfactorily whilst performing efficiently under automatic reasoning. Consideration should also be given to its potential reuse and extensibility. In practice, the interplay between all these factors may lead to a number of compromises.

Protégé [44] is a convenient, free, and open-source ontology-editing tool that provides a friendly user interface for creating and testing axioms. Such editing tools are particularly useful for aiding the design process in which the most appropriate design patterns may not be immediately obvious. Taking the example of the ICD-O-3 [45] spindle cell sarcoma with morphology code 8801 and tumor behavior code 3 (malignant behavior), the compound code (morphology-behavior) can be modeled in the ontology in several ways (where the morphology code has been prepended with the letter "M_" for more convenient class-naming purposes):

$$M\_8801\_3 \sqsubseteq M\_8801 \sqcap BehaviorCode3 \tag{1}$$

$$M\_8801\_3 \sqsubseteq M\_8801 \sqcap \exists hasBehaviour.BehaviorCode3 \tag{2}$$

$$M\_8801 \sqcap BehaviorCode3 \sqsubseteq M\_8801\_3 \tag{3}$$

Eqs. (1) and (2) are similar apart from the fact that behavior in Eq. (2) has been expressed in terms of an existential restriction. Behavior may not even need to be modeled at all and just left implicit in the name of the class (since the trailing digit denotes the behavior code). The choice ultimately depends on how the morphology-behavior class will be used in other classes. For instance, a prostate tumor can have ICD-O-3 topography code C619, morphology code 8801, and behavior code 3 and may be modeled in a similar fashion to Eqs. (1)–(3):

$$ProstateTumor \sqsubseteq C619 \sqcap M\_8801\_3 \tag{4}$$

$$ProstateTumor \sqsubseteq \exists hasTopography.C619 \sqcap \exists hasMorphology.M\_8801\_3 \tag{5}$$

$$C619 \sqcap M\_8801\_3 \sqsubseteq ProstateTumor \tag{6}$$

It could also be modeled as an Abox axiom to denote that this is a specific instance of a more general prostate cancer class. It is not necessarily a simple choice since there are advantages and disadvantages to each approach. With Eq. (5) the concepts of topography and morphology can be declared disjoint (a topography is not a morphology), but then modeling a tumor type or signature (e.g. $\exists hasTumorSignature.ProstateTumor$) would hide the topography and morphology codes in two existential restrictions:

$$\exists hasTumorSignature.(\exists hasTopography.C619 \sqcap \exists hasMorphology.M\_8801\_3) \tag{7}$$

and thereby makes it a harder task to access the code values without increasing the language expressivity (such as including inverse operations or complex role inclusion axioms or other rules). It would be even harder to access the behavior code had Eq. (2) been used owing to the chain of existential restriction. Eq. (6) results in automatic class subsumption of the conjunction $C619 \sqcap M\_8801\_3$ under the class *ProstateTumor* but can lead to higher processing costs than Eq. (4) [46].

Nevertheless, subsumption is a primary mechanism used by automatic reasoners to make inferences on a knowledge base and is perhaps the most critical factor to take into account in the design of an ontology that models validation rules predominantly using TBox axioms. OWL uses the open world assumption (OWA) in which the truth of a statement is unknown unless it is expressly known to be true/false – the philosophy being that there may always be extra information not yet declared in the knowledge base that has further bearing on the statement. The consequence is that an entity having topography C619 and morphology M_8801_3 would not be considered as a *ProstateTumor* using Eq. (4) for the reason that there may be other as-yet undisclosed information to describe it further. The work-around would be either to make an equivalence – which can lead to subtle unintended consequences in more complex expressions – or to use the form of Eq. (6), which Protégé refers to as a general concept inclusion (GCI). CGIs provide several benefits in the correct context [47].

Also relevant is the balance between pre- and post-coordination of the ontology [48] – in pre-coordination, all the relationships are explicitly declared a priori, whereas in post-coordination a reasoner is used to infer relationships between entities a posteriori. In addition, other types of rules can be incorporated into OWL ontologies using the Semantic Web Rule Language (SWRL). SWRL extends the expressivity of OWL DLs using Horn-like logic rules (in which logic statements are written in terms of an implication) and can overcome some limiting cases in OWL at the potential cost of decidability and interoperability [49]. **Table 4** summarizes some of the more important mechanisms that can be employed in validation-type tests.

There are thus a number of careful choices to be made dependent upon how the ontology will be used. The consequence of these design decisions may compromise the ability to reuse existing ontologies as well as render the ontology developed unsuitable for wider purposes.

3.3 Data shapes languages

An alternative to using an ontology for data validation, but which still draws directly from the data model, is to use a data shapes language such as the Shapes

Pre/post coord	Mechanism	Utilization	Advantages/disadvantages
Post	Subsumption	Defined classes (TBox)	Ensures subsumption (since classes are equivalent). Can give rise to unintended equivalences
Post	Subsumption	General Concept Inclusions (TBox)	Ensures subsumption if the ontology design is correct. Needs careful ontology design to ensure the specific order of subsumption, which may conflict with other requirements
Post	Subsumption	Individuals and higher DL expressivities (ABox)	Greater flexibility and functionality. More difficult to control logic, and computationally expensive
Post	Inconsistency of class structure	Disjoint class definitions	Straightforward to catch any validation errors. Can lead to unintended class inconsistencies for ontologies with many class inter-relations
Post	Additional logic (internal to ontology)	SWRL	Provides extra functionality. Difficult to control if many rules and can lead to portability issues
Both	Additional logic (external to ontology)	Programming logic	Considerable control and extra functionality. Requires a dedicated computer program and extra maintenance
Pre	Comprehensive assertions	Predefinition of all entities and relationships	All the relationships are known a priori. Ontology can be very large and lead to performance issues if interfaced with ontologies requiring automatic reasoning

Table 4.
Summary of the most important ontology-based mechanisms that can used for data validation purposes with their main associated advantages/disadvantages.
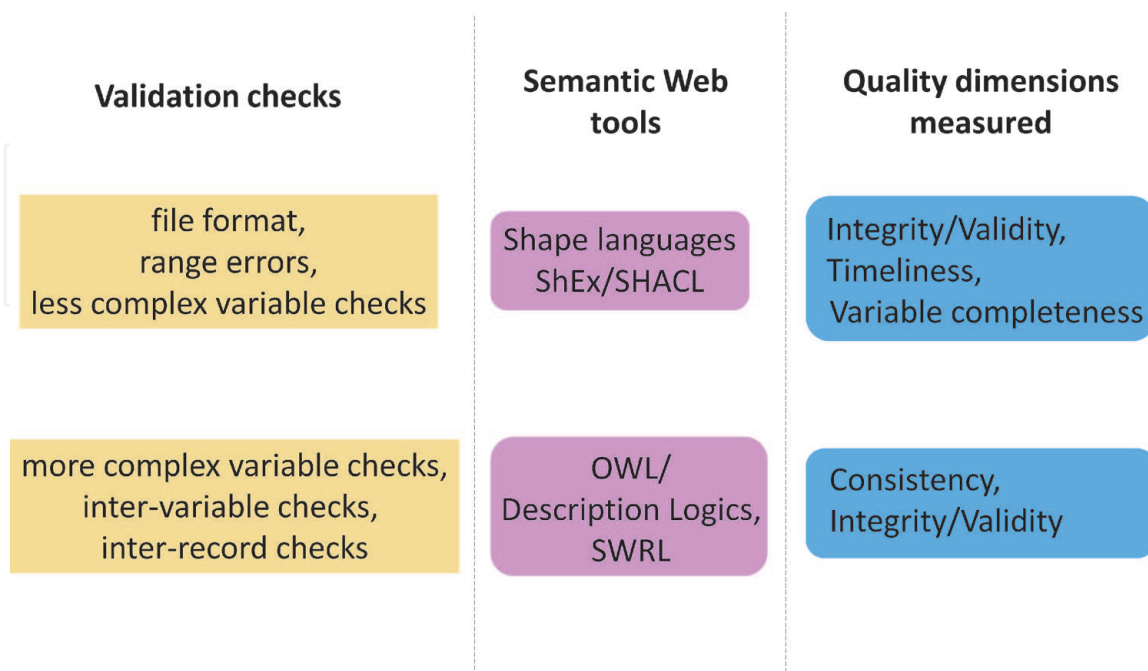

Figure 2.
Applicability of the semantic-web tools to the different steps of the validation process and the quality dimensions they are able to measure. Shape languages such as ShEx and SHACL provide the means for finding non-compliance to the more straightforward data validation rules. More complex validation checks require the increased functionality offered by DLs maybe in combination with SWRL and dedicated program logic.

Constraint Language (SHACL) [50] or Shapes Expressions (ShEx) [51]. Both languages benefit from the possibility of formulating the rules under the closed world assumption (CWA) which, contrary to the OWA, considers a statement to be false unless it has otherwise explicitly been declared to be true.

The degree of complexity that can be handled for the inter-variable validation checks is more limited, but in cases where this does not pose a problem, SHACL in particular provides a number of advantages. SHACL is specifically intended as a language for describing constraints on RDF data and has been used to describe ontology design patterns for validating data in Electronic Health Records (EHRs) conforming to Clinical Information Modeling Initiative (CIMI) models [52].

ShEx can also be used to validate data but the underlying philosophy is different from that of SHACL. As noted in [53], ShEx is more grammar-related whilst SHACL is more constraint-related with the result that ShEx puts greater focus on validation results in contrast to SHACL that gives more attention to validation errors. As discussed in Section 4.4, ShEx is particularly useful in detecting syntactic and range errors in the preprocessing stages of CR data validation. **Figure 2** provides an overview of the applicability of the semantic-web tools to the different data-validation steps and the quality dimensions they are able to address.

4. Quality criteria for CR data

Before CR data can be compared at inter-regional or international level, they have to pass through a rigorous cleaning process. From the point of view statistical analysis, assessment of the quality and reliability of data hinges on the basic requirement of the representativeness of the data. A large CR data set for which reasonable doubts exist concerning the data representativeness has less value than a small CR data set with high representativeness.

More specifically for statistical analyses to derive incidence and survival indicators from CR data, the two required dimensions are completeness (the confidence that all diagnosed cancers in the population are actually included in the data set) and accuracy (the confidence that the proportion of cases with a given set of characteristics truly reflects reality [37, 54]). Whereas timeliness is another important dimension [54], it may lead to some trade-off with the degree of data completeness [55].

One cause of incompleteness observed in cancer survival studies results from the varying risk of death from other causes than cancer, and is more pronounced for the older age brackets [56] (competing-risks phenomenon). Other observational studies performed with the availability of additional, post factum data reveal that the level of incompleteness can also be cancer-site specific [55].

In addition, high-quality cancer data should have high comparability between different populations over time, which can best be achieved using up-to-date, homogeneous, and consistent data collection and recording procedures [54]. Application of the standard data validation rules is one way of ascertaining the comparability of data between different CRs, as discussed in the following sub-sections.

4.1 Inferring TNM stage

TNM (Tumor, Nodes, Metastases) is a globally recognized cancer staging classification system for describing the extent and spread of solid tumors in terms of tumor size, invasion of lymph nodes, and presence of metastases. One of the validation checks relates to the validity of TNM stage on the basis of the associated TNM parameters (including: topography, morphology, pathological/clinical T, N, and M codes, TNM edition, as well as age, and grade for certain tumor sites).

Validity can be ascertained using the automatic reasoner to infer the stage from the parameters and compare it with the value provided by the registry. Axioms to model stage can be defined along the lines of the example taken for stage I prostate cancer:

$$TNMEd7SiteProstate \sqcap \exists hasBehavior.BehaviorCode3 \sqcap \exists hasT.(T1 \sqcup T2a) \sqcap \\ \exists hasN.N0 \sqcap \exists hasM.M0 \sqsubseteq TNMStageI \tag{8}$$

in which:

$$\exists hasTopography.C619 \sqcap \exists hasMorphology.Carcinoma \sqsubseteq TNMSiteProstate \tag{9}$$

$$TNMEd7SiteProstate \sqcap \exists hasTNMEdition.TNMEd7 \sqsubseteq TNMEd7SiteProstate \tag{10}$$

and all the ICD-O-3 morphologies associated with carcinoma have the form similar to:

$$\exists hasMorphology.M\_8140 \sqsubseteq \exists hasMorphology.Adenocarcinoma \tag{11}$$

in which, for example:

$$Adenocarcinoma \sqsubseteq Carcinoma \tag{12}$$

The resulting subsumption process for a CR case record passed in with the values: topography C619, morphology 8140, TNM edition 7, and TNM parameters: T2a, N0, M0 would be the following:

a. morphology M_8140 is subsumed under the class *Carcinoma* from Eqs. (11) and (12);

b. topography C619 together with the subsumed morphology M_8140 under the class *Carcinoma*, are further subsumed under the class *TNMEd7SiteProstate* from Eqs. (9) and (10);

c. the subsumpton result of (b) together with the specified TNM parameters, are finally subsumed under the stage class *TNMStageI*.

The value of stage inferred by the reasoner can then be compared with the stage value provided with the CR case record in order to validate the record. Axioms described in this manner can be developed to provide a modular structure to model TNM stage for all editions of TNM.

4.2 Multiple primary tumors validation check

For the purpose of deriving cancer incidence indicators, it is important in patients with multiple cancer case records to distinguish between tumors that are linked with an existing case and those that are not. The latter are referred to as multiple primary tumors and they need to be validated.

An international set of rules provides the definition of multiple primary tumors [57]. Transcribing the rules into DL requires a higher expressivity owing to the need for ABox statements, inverse relationships, and qualified number restrictions. These requirements arise from the need to analyze the different permutations of the possible tumor pairings according to the rules. The latter can be transcribed as a set of TBox axioms which are used by the reasoner to test the dependencies of multiple

tumor cases defined as a set of ABox axioms. TBox axioms take the form of constructs encapsulated in Eqs. (13)–(16) below (described in greater detail in [58]):

$$\exists\, hasMorphology.MorphGroupX \sqcap \exists\, hasMorphoplogy.MorphGroupXDep$$
$$\sqsubseteq DuplicateMorphologyGroup \tag{13}$$

Eq. (13) models the conjunction of two dependent morphology groups as a sub-class of the class depicting a duplicate morphology, according to one of the multiple primary tumor rules:

$$DuplicateMorphologyGroup \sqcap \exists\, hasMorphoplogy.ICDO3HematologicalMorphology$$
$$\sqsubseteq DuplicatePrimaryCondition \tag{14}$$

Eq. (14) models the conjunction of a previously-determined duplicate morphology with a hematological morphology type as a duplicate primary tumor condition, according to another of the multiple primary tumor rules.

$$\geq 2 hasTopography.(C26 \sqcup C68 \sqcup C76) \sqsubseteq DuplicateTopographyGroup \tag{15}$$

Eq. (15) models the rule that if the two topographies of a tumor pairing are in any of the "other or ill-defined" topography groups or subgroups they are considered a duplicate topography group.

$$DuplicateMorphologyGroup \sqcap DuplicateTopographyGroup \sqsubseteq$$
$$DuplicatePrimaryCondition \tag{16}$$

Eq. (16) models a resulting duplicate primary tumor for the case of a duplicate morphology and a duplicate topography.

ABox axioms are built up using permutations of tumor morphologies and topographies, where a tumor is defined by the TBox axiom as the conjunction of one morphology and one topography:

$$ICDO3Tumor \equiv = 1\, hasMorphology.ICDO3Morphology \sqcap$$
$$= 1\, hasTopography.ICDO3Topography \tag{17}$$

Accessing the morphologies from two tumor individuals to derive a morphology permutation can be performed using the ABox axiom:

$$p1\_tpM1 : \exists\, hasMorphology.(\exists\, hasMorphology^{-}.(p1\_t1)) \sqcap$$
$$\exists\, hasMorphology.(\exists\, hasMorphology^{-}.(p1\_t2)) \tag{18}$$

where the name of the individual $p1\_tpM1$ refers to the first morphology permutation of the two individual tumors $p1\_t1$, and $p1\_t2$ of a given patient $p1$. Since the morphologies have already been assigned in the tumor ABox axioms according to the pattern of Eq. (17), their specific values can be extracted using the inverse relationships in Eq. (18). Similar axioms can be defined for the topography permutations.

ABox axioms for the tumor pairings (containing a morphology pairing and a topography pairing) can then be specified according to the template:

$$(p1\_tc1, p1\_tpM1) : \exists\, hasTumorPermutationMorphology \sqcap$$
$$(p1\_tc1, p1\_tpT1) : \exists\, hasTumorPermutationTopography \tag{19}$$

in which $p1\_tc1$ refers to the first tumor pairing for patient $p1$ and $p1\_tpM1$ and $p1\_tpT1$ refer to the first morphology pair and topography pair respectively.

The axioms can be constructed automatically from the input records since the cancer-case records have a patient identifier and a tumor identifier and therefore all the tumor-pairing permutations can be ascertained in a preprocessing step. On the basis of the TBox axioms, the reasoner classifies the ABox axioms under the class *DuplicatePrimaryCondition* for instances where the multiple-primary rules are violated.

4.3 Tumor signature validation check

The third batch of validation checks concerns the specificities of tumor types, particularly in relation to parameters including basis of diagnosis, grade, age at diagnosis, sex, and topography-morphology-behavior inter-dependencies. These checks concern many of the rule tables provided in [15] and are examples of rules that be modeled in a variety of ways as discussed in Section 3.2 and which ultimately can be related to the balance between pre- and post-coordination of classes [48].

A tumor type, which we refer to as a tumor signature, comprises a topography/ set of topographies in association with a set of morphologies. The topographies and morphologies may additionally specify a number of restrictions on values of associated variables such as age of patient at diagnosis, sex, basis of diagnosis, grade, etc.

Pre-coordination allows the greatest control over the definition of tumor signatures since it allows each tumor signature at its most granular level to be defined independently. Consequently, the permissible ranges of values of all the dependent variables can be specified for each tumor signature individually. The drawback to this approach is that it would result in over 200,000 unique tumor-signature classes and could have implications on reasoning speeds of other ontologies that use them.

The design used by SNOMED CT [59] to handle all the possible clinical terminology class definitions is to create a number of general classes in a pre-coordinated way and capture the specializations of those classes either in equivalent classes or GCI expressions that would be determined in a post-coordinated way by means of the reasoner [48]. Emulating such a design would allow, for instance, the qualifying rule of age on a given morphology/set of morphologies to be expressed as a specialization. Taking as an example the morphology M_8970 (Hepablastoma) which has a qualifying rule for ages greater than five, the associated morphology class can be sub-classed from a data property such that:

$$M\_8970 \sqsubseteq \exists\, ageAtDiagnosis.\{\geq 6\} \qquad (20)$$

The resulting subsumption for hepablastomas thereby provides a mechanism through which it can be ensured that all qualifying rules are respected in the data.

4.4 Data quality metric

Once the rules have been established in the ontology, the individual data records can be validated according to the various groups of tests (e.g. stage, multiple primaries, tumor signature, etc.). Of the seven generally agreed dimensions for data quality listed in Section 2.2, integrity, consistency, and variable-completeness of CR common data sets are ascertained in a relatively straightforward manner for each of the tests and scored in percentage terms of conforming records using the metric:

$$\left(1 - \frac{R_e}{R_T}\right) \times 100 \qquad (21)$$

where R_e is the total number of non-conforming records to the particular test parameters and R_T is the total number of records used within the test. Eq. (21) takes a similar form to that proposed in [60] for both completeness and consistency quality dimensions and was assessed in [38] to fulfill all the five data-metric requirements discussed in Section 2.2.1.

Variable-completeness would describe the extent of the availability of information/variables necessary for running the specific test. Integrity would provide information on the number of records passing the test. Consistency would then be a measure of data conformity across tests – e.g. consistency of the morphology-topography code combinations not just within one individual test but across all tests (TNM, multiple primary and tumor signature).

The syntactic part of the integrity dimension (as differentiated in ISO 8000-8) can be measured from a preprocessing stage which in general is necessary to ensure the correct format of the cancer-case records before passing them into the DL-based validation checks. This preprocessing stage can itself be performed also with direct reference to the data model using a shape language such as ShEx as discussed in Section 3.3. ShEx is particularly appropriate for validating the format and ranges of the variable values and benefits from the possibility of formulating the rules under the closed world assumption. The output of this stage can therefore provide a metric for data-type integrity also in percentage terms of records conforming to the ShEx schema.

As noted in Section 2.2.1, the quality dimensions posing greatest difficulty are data-completeness and accuracy. Various metrics to estimate the former have been proposed [37] and those based on mortality-incidence ratios or survival probabilities conform well to the data-metric requirements of [38]. Whereas accuracy issues may be insinuated from the result of the integrity/validity checks, the surest way of detecting them would be through a data-auditing process such as that advocated by ISO 8000-8.

4.5 Process automation

The chain of processes from preparation of the CR common data set to reading cancer case records into the ontology and performing the validation checks and counting the non-conforming records can be automated using the OWL application program interface (OWL-API) [61]. The OWL-API provides methods for accessing the ontology axioms, invoking the reasoner, and polling the results of the reasoning process. The API also allows the incorporation of program logic to permit greater expressivity although at the expense of increased maintenance.

The strength of the ontological approach is that the data model and the data-quality model – at least for the integrity and consistency dimensions remain in synchronization owing to the fact that they are integrated in the same sets of ontologies. Not only does this aid transparency of the validation process but it also simplifies maintenance and version control via the URIs pointing to the most current version of the ontology.

Moreover, the outputs of the validation process are readily verifiable by a trusted third party since it would basically be a matter of rerunning the checks on the CR file and comparing the outputs. For situations where the integrity of the quality metrics is important, the trusted third party can provide such assurance by integrating the validation checks together with the tests for data completeness and accuracy into a data-quality certification scheme such as ISO 8000-8.

5. Constructing a data-quality context

The quality context is as important as the semantic context for interoperability of CR data and as applicable to machine-based reasoning as it is to human-based

reasoning; even though the semantics might admit the apparent compatibility of data sets, any inferences drawn from their combination could be legitimately challenged without due attention to the data quality. The importance of taking CR data completeness into consideration when comparing survival estimates between different populations has been emphasized previously [62]. In short, data quality is a critical issue for health data where erroneous inferences could lead to potentially dire consequences [63]. Encapsulating quality metrics in the metadata associated with the data set would adapt well to the FAIR digital object framework, and indeed such a model was proposed as far back as 1999 [64] and more recently in [65].

Agreeing a common set of data-quality metrics is however not an easy task and perhaps explains the lack of an overall framework. Whereas the difficulties are more acute for unstructured data [66] and require complicated semantic enrichment techniques [66], processes dealing with structured data pose less difficulty. The key to a potentially elegant solution able to unify both semantic and quality aspects of interoperability may lie in the use of OWL ontologies for describing common data models, or at least relevant parts of them.

If designed carefully, OWL axioms can be used for validating CR data sets against predefined rules as discussed in Section 4, thereby providing a quantitative quality index or set of indices for certain quality dimensions on which to base pragmatic decisions regarding the compatibility/comparability of different data sets. The availability of such a decision framework is critical to any eventual devolution of the centralized data-cleaning processes to the local level. It is also critical for purposes of secondary-data usage where the end user/application has to be aware of issues limiting the extent and purpose for which different data sets can be used.

With respect to the generally agreed seven quality dimensions, completeness of the mandatory variable set (variable-completeness), integrity and consistency are
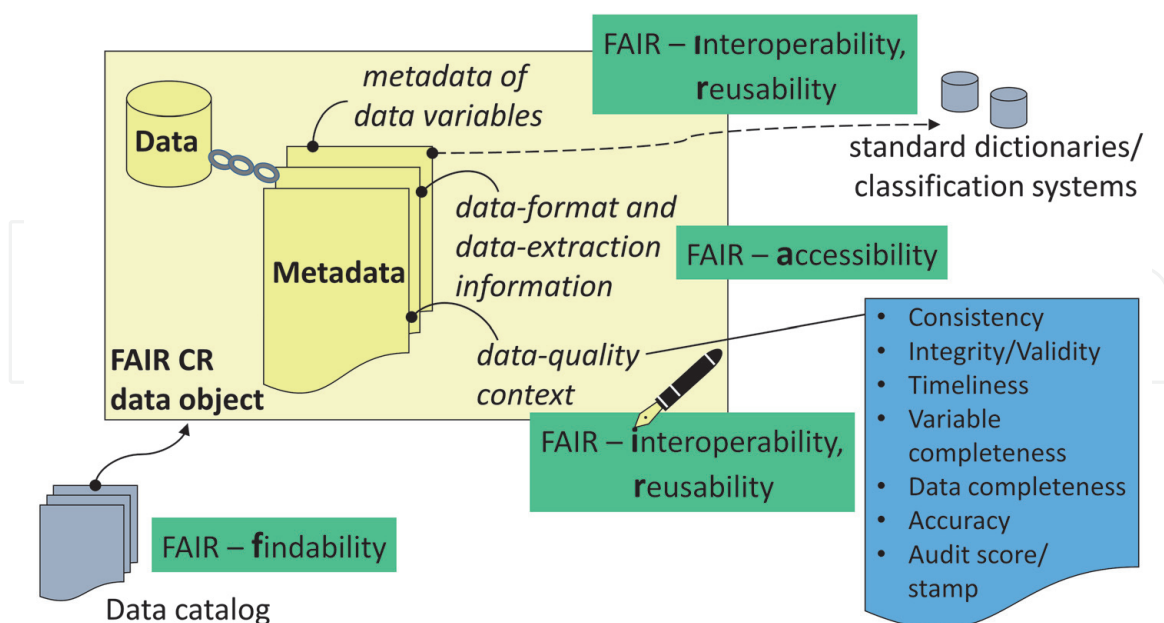

Figure 3.
Depiction of a FAIR cancer-registry data set in terms of a FAIR digital object (FDO). The FDO comprises the data itself and an associated set of metadata components that describe the data and their context. The FDO is registered in a catalog to make it findable. One of the metadata components provides information on how to access the data. Another metadata component describes the metadata and semantics of the data-set variables and links to standard dictionaries using the semantic relations of knowledge organization systems (e.g. SKOS). The semantic context provides an essential part of data interoperability and reusability. A further metadata component provides the data-quality context and "dots the 'i' of interoperability" by adding the second vital ingredient towards making the data interoperable and reusable.

ascertainable from the validation process of the CR common data sets with each dimension being measured in percentage terms of conforming records as suggested in [35] and according to Eq. (21). Uniqueness can be ensured by a correct definition of the common data set template and therefore be provided as a default measure for all CR common data sets. Timeliness can be determined directly from the data set variable relating to cancer-case registration date providing a metric easy to measure. Data-completeness can be estimated in several ways as discussed in [37], one of which also provides a quantifiable metric along the lines of Eq. (21). The metrics for these quality dimensions would therefore all fulfill the requirements stipulated for a data metric supporting a decision-based framework [38].

The remaining quality dimension, accuracy, is dependent on the primary-data capture process, which is outside the control of cancer registries. Whereas, performance of the validation checks and frequency analyses of selected variables may provide some proxy measures for systematic errors, a more robust method would need a data-auditing process in the various stages of the data pipeline. The resulting accuracy metrics could then be passed along through each stage to form a compound accuracy measure on the data set.

In this way, a comprehensive and structured data-quality context could be constructed and thereafter provided as an additional component of the associated FAIR digital object, as illustrated in **Figure 3**. This component would provide a direct means for decision-based mechanisms to compute quantitative differences between quality measures of data sets and thereby infer the suitability of their integration in some fashion.

6. Conclusions

Achieving data interoperability, at least in the widest sense, is a major challenge. In order to be able to integrate or compare heterogeneous data sets, data users non-expert in the respective data domains need a considerable amount of contextual information. Whereas these needs can be met partially by semantic linkage of metadata, the aspect of data quality is crucial especially in quality-critical disciplines such as health. The FAIR data principles acknowledge the importance of data quality but do not address it directly.

A means of quantifying the data quality context in CR data sets along a number of representative and widely accepted quality dimensions has been presented. These metrics provide a quality context that can serve as an additional set of metadata within the associated FAIR digital object and made available with any aggregated data derived from it. The latter is an important consideration for entities having access only to the aggregated data sets for which the information is no longer available to verify the data quality directly from the validation rules themselves.

Having access to this type of data-quality information, even if measured in relatively simple terms, would enable data-processing entities to make certain informed decisions on the likely compatibility with other data sets. Not only is this a fundamental prerequisite to being able ultimately to federate the CR data-harmonization processes themselves but also to promoting the availability of CR data in ways that would prove useful and informative for secondary-data purposes. It would also allow more scrutiny and transparency on the results of secondary analyses that may have potentially far-reaching consequences.

Although the focus has been on CR data, the ideas are sufficiently generic to apply as a general framework to other data domains and is amenable to formalization in a data-quality auditing process such as ISO 8000-8 by providing a conceptual model and the defined means of verification against the model.

Acknowledgements

Conflict of interest

The authors declare that they have no competing interests.

Author details

Nicholas Nicholson*, Francesco Giusti, Luciana Neamtiu, Giorgia Randi, Tadeusz Dyba, Manola Bettio, Raquel Negrao Carvalho, Nadya Dimitrova, Manuela Flego and Carmen Martos
European Commission Joint Research Centre, Ispra, Italy

*Address all correspondence to: nicholas.nicholson@ec.europa.eu

IntechOpen

References

[1] Parkin DM. The evolution of the population-based cancer registry. Nature Reviews. Cancer. 2006;**6**: 603-612. DOI: 10.1038/nrc1948

[2] Parkin DM. The role of cancer registries in cancer control. International Journal of Clinical Oncology. 2008;**13**:102-111. DOI: 10.1007/s10147-008-0762-6

[3] dos Santos Silva I. Cancer Epidemiology: Principles and Methods, Ch 17. Lyon, France: IARC Press; 1999. 442 p. Available from: https://publications.iarc.fr/Non-Series-Publications/Other-Non-Series-Publications/Cancer-Epidemiology-Principles-And-Methods-1999

[4] Bray F, Znaor A, Cueva P, et al. Planning and Developing Population-Based Cancer Registration in Low- and Middle-Income Settings. 2014. Available from: https://www.who.int/immunization/hpv/iarc_technical_report_no43.pdf [Accessed: July 26, 2021]

[5] Public Health Scotland. Scottish Cancer Registry – How Data are Collected. Available from: https://www.isdscotland.org/Health-Topics/Cancer/Scottish-Cancer-Registry/How-data-are-collected/ [Accessed: July 26, 2021]

[6] Tucker TC, Durbin EB, McDowell JK, Huang B. Unlocking the potential of population-based cancer registries. Cancer. 2019;**125**:3729-3737. DOI: 10.1002/cncr.32355

[7] Thompson CA, Jin A, Luft HS, Lichtensztajn DY, Allen L, Liang SY, et al. Population-based registry linkages to improve validity of electronic health record-based cancer research. Cancer Epidemiology, Biomarkers & Prevention. 2020;**29**(4):796-806. DOI: 10.1158/1055-9965.EPI-19-0882

[8] NIH Eunice Kennedy Shriver National Institute of Child Health and Human Development. Data Harmonization. Available from: https://www.icpsr.umich.edu/icpsrweb/content/DSDR/harmonization.html [Accessed: July 26, 2021]

[9] Arndt V, Holleczek B, Kajüter H, Luttmann S, Nennecke A, Zeissig SR, et al. Data from population-based cancer registration for secondary data analysis: Methodological challenges and perspectives. Das Gesundheitswesen. 2020;**82**(Suppl. 1):S62-S71. DOI: 10.25646/6907

[10] Antonio AS, Ferlay J, Soerjomataram I, Znaor A, Jemal A, Bray F. Bladder cancer incidence and mortality: A global overview and recent trends. European Urology. 2017;**71**(1): 96-108

[11] National Cancer Institute. North American Surveillance, Epidemiology, and End Results (SEER) Program. Available from: https://seer.cancer.gov/ [Accessed: July 26, 2021]

[12] European Commission. European Cancer Information System (ECIS). Available from: https://ecis.jrc.ec.europa.eu/ [Accessed: July 26, 2021]

[13] International Agency for Research on Cancer. Cancer Incidence in Five Continents (CI5). Available from: https://ci5.iarc.fr/Default.aspx [Accessed: July 26, 2021]

[14] International Agency for Research on Cancer. Global Cancer Observatory. Available from: https://gco.iarc.fr/ [Accessed: July 26, 2021]

[15] Martos C, Crocetti E, Visser O, Rous B, Giusti F, et al. A proposal on cancer data quality checks: one common procedure for European cancer registries. JRC Technical Report, version 1.1. Luxembourg: Publications office of

the European Union; 2018. 99 p. DOI: 10.2760/429053

[16] Wilkinson M, Dumontier M, Aalbersberg I, et al. The FAIR guiding principles for scientific data management and stewardship. Scientific Data. 2016;**3**: 160018. DOI: 10.1038/sdata.2016.18

[17] European Commission Expert Group on FAIR Data. European Commission Directorate General for Research and Innovation. Turning FAIR Into Reality. Luxembourg: Publications office of the European Union; 2018. 76 p. DOI: 10.2777/1524

[18] IDC. The Secondary Use of Health Data and Data-driven Innovation in the European Healthcare Industry. 2020. Available from: https://datalandscape.eu/sites/default/files/report/D3.6_Data-driven_Innovation_in_Health_21.01.2020_Final.pdf [Accessed: July 26, 2021]

[19] European Commission. European Health Data Space. Available from: ec. europa.eu/health/ehealth/dataspace_en [Accessed: July 26, 2021]

[20] W3C. Data Catalog Vocabulary (DCAT) – Version 2 Recommendation. 2020. Available from: https://www.w3. org/TR/vocab-dcat-2/ [Accessed: July 9, 2021]

[21] ISO/IEC. Information Technology – Metadata Registries (MDR). Part 1: Framework. 2015. Available from: https://www.iso.org/standard/61932. html [Accessed: July 9, 2021]

[22] W3C. SKOS Simple Knowledge Organization System. Available from: https://www.w3.org/2004/02/skos/ [Accessed: July 9, 2021]

[23] Fiume M, Cupak M, Keenan S, et al. Federated discovery and sharing of genomic data using Beacons. Nature Biotechnology. 2019;**37**:220-224. DOI: 10.1038/s41587-019-0046-x10.1038/s41587-019-0046-x

[24] Global Alliance for Genomics and Health. GA4GH Genome Beacons. Available from: https://beacon-project. io/categories/howto.html [Accessed: July 9, 2021]

[25] Sinaci AA, Laleci Erturkmen GB. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. Journal of Biomedical Informatics. 2013;**46**:784-794. DOI: 10.1016/j.jbi.2013.05.009

[26] Nicholson N, Perego A. Interoperability of population-based patient registries. Journal of Biomedical Informatics. 2020;**112s**:100074. DOI: 10.1016/j.yjbinx.2020.100074

[27] MOLGENIS Data Platform. Available from: https://www.molgenis. org/ [Accessed: July 9, 2021]

[28] Apache Atlas. Available from: https://atlas.apache.org/#/ [Accessed: July 9, 2021]

[29] Corcho O, Eriksson M, Kurowski K, Ojsteršek M, Choirat C, van de Sanden Mark, Coppens F. EOSC interoperability framework - Report from the EOSC Executive Board Working Groups FAIR and Architecture. Luxembourg: Publications office of the European Union; 2021. 60 p. DOI:10.2777/620649

[30] Bonino da Silva Santos LO. FAIR Digital Object Framework Documentation Working Draft; Leiden: GO FAIR Foundation; 2021. Available from: https://fairdigitalobjectframe work.org/ [Accessed: July 9, 2021]

[31] De Smedt K, Koureas D, Wittenburg P. FAIR digital objects for science: From data pieces to actionable knowledge units. Publica. 2020;**8**(2):21. DOI: 10.3390/publications8020021

[32] Data Interoperability Standards Consortium. Available from: https:// datainteroperability.org/ [Accessed: July 9, 2021]

[33] GO FAIR. What FAIR is Not … . Available from: https://www.go-fair. org/resources/faq/what-fair-is-not/ [Accessed: July 9, 2021]

[34] Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. Data Science Journal. 2015;**14**:2. DOI: 10.5334/dsj-2015-002

[35] DAMA UK. The Six Primary Dimensions For Data Quality Assessment. Bristol: DAMA UK; 2013. Available from: https://docplayer.net/ 3987248-The-six-primary-dimensions-for-data-quality-assessment.html [Accessed: July 9, 2021]

[36] ISO. Data Quality – Part 8: Information and Data Quality: Concepts and Measuring ISO 8000-8. Geneva, Switzerland: ISO; 2015

[37] Parkin DM, Bray F. Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness. European Journal of Cancer. 2009;**45**(5):756-764

[38] Heinrich B, Hristova D, Klier M, Schiller A, Szubartowicz M. Requirements for data quality metrics. Journal of Data and Information Quality. 2018;**9**(2):1-32. DOI: 10.1145/3148238

[39] National Center for Biomedical Ontology. Bioportal. Available from: https://bioportal.bioontology.org/ [Accessed: July 9, 2021]

[40] W3C. Web Ontology Language (OWL). Available from: https://www. w3.org/OWL/ [Accessed: July 9, 2021]

[41] W3C. Resource Description Framework (RDF). Available from: https://www.w3.org/RDF/ [Accessed: July 9, 2021]

[42] Knorr M, Hitzler P. Description logics. In: Siekmann JH, editor. Handbook of the History of Logic. Vol. 9. The Netherlands: Elsevier

Radarweg, AE Amsterdam; The Netherlands; 2014. pp. 659-678. DOI: 10.1016/B978-0-444-51624-4.50015-0

[43] Baader F, Horrocks I, Lutz C, Sattler U. An Introduction to Description Logic, Ch 1. Cambridge, UK: Cambridge University Press; 2017

[44] Protégé. A Free, Open-Source Ontology Editor and Framework for Building Intelligent Systems. Available from: https://protege.stanford.edu/ [Accessed: July 9, 2021]

[45] World Health Organization. International Classification of Diseases for Oncology (ICD-O) – 3rd Edition, 1st Revision. 2013. Available from: https://apps.who.int/iris/handle/10665/ 96612 [Accessed: July 26, 2021]

[46] Hammar K. Reasoning performance indicators for ontology design patterns. In: Proceedings of the 4th International Conference on Ontology and Semantic Web Patterns (WOP'13); Aachen, Germany: CEUR-WS; 2013. pp. 27–38

[47] Sattler U, Stevens R. Being complex on the left-hand-side: General Concept Inclusions. Ontogenesis. 2012. Available from: http://ontogenesis.knowledgeblog. org/1288 [Accessed: July 9, 2021]

[48] Stevens R, Sattler U. Post-coordination: Making things up as you go along. Ontogenesis. 2013. Available from: http://ontogenesis.knowledgeblog. org/1305 [Accessed: July 9, 2021]

[49] W3C. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. Available from: https://www.w3.org/Sub mission/SWRL/ [Accessed: July 9, 2021]

[50] W3C. Shapes Constraint Language (SHACL). Available from: https://www. w3.org/TR/shacl/ [Accessed: July 9, 2021]

[51] W3C. Shape Expressions Language (ShEx). Available from: http://shex.io/she x-semantics/ [Accessed: July 9, 2021]

[52] Martínez-Costa C, Schulz S. Validating EHR clinical models using ontology patterns. Journal of Biomedical Informatics. 2017;**76**:124-137. DOI: 10.1016/j.jbi.2017.11.001

[53] Labra Gayo JE, Prud'hommeaux E, Boneva I, Kontokostas D. Validating RDF Data. In: Ding Y, Groth P, series editors. Synthesis Lectures on Semantic Web: Theory and Technology, Lecture #16. San Rafael, California, USA: Morgan & Claypool Publishers; 2018. 304 p. DOI: 10.2200/ S00786ED1V01Y201707WBE016

[54] Bray F, Parkin DM. Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness. European Journal of Cancer. 2009; **45**(5):747-755

[55] Zanetti R, Schmidtmann I, Sacchetto L, Binder-Foucard F, Bordoni A, Coza D, et al. Completeness and timeliness: Cancer registries could/ should improve their performance. European Journal of Cancer. 2015;**51**(9): 1091-1098

[56] Schuster NA, Hoogendijk EO, Kok AAL, Twisk JWR, Heymans MW. Ignoring competing events in the analysis of survival data may lead to biased results: A nonmathematical illustration of competing risk analysis. Journal of Clinical Epidemiology. 2020;**122**:42-48. DOI: 10.1016/j.jclinepi.2020.03.004

[57] International Association of Cancer Registries. International rules for multiple primary cancers. Asian Pacific Journal of Cancer Prevention. 2005; **6**(1):104-106

[58] Nicholson NC, Giusti F, Bettio M, Negrao Carvalho R, Dimitrova N, Dyba T, et al. An ontology to model the international rules for multiple primary malignant tumours in cancer registration. Applied Sciences. 2021;**11**: 7233. DOI: 10.3390/app11167233

[59] SNOMED CT. Available from: http:// www.snomed.org [Accessed: July 9, 2021]

[60] Blake R, Mangiameli P. The effects and interactions of data quality and problem complexity on classification. Journal of Data and Information Quality. 2011;**2**(2):1-28. DOI: 10.1145/ 1891879.1891881

[61] Horridge M, Bechhofer S. The OWL API: A java API for OWL ontologies. Semantic Web. 2011;**2**(1):11-21. DOI: 10.3233/SW-2011-0025

[62] Robinson D, Sankila R, Hakulinen T, Moller H. Interpreting international comparisons of cancer survival: The effects of incomplete registration and the presence of death certificate only cases on survival estimates. European Journal of Cancer. 2007;**43**:909-913

[63] The Connecting for Health Common Framework. Background Issues on Data Quality. 2006. Available from: http://bok.ahima.org/PdfView? oid=63654 [Accessed: July 14, 2021]

[64] Vassiliadis P, Bouzeghoub M, Quix C. Towards Quality-Oriented Data Warehouse Usage and Evolution. In: Advanced Information Systems Engineering, 11th International Conference (CAiSE'99); 14-18 June 1999; Berlin, Heidelberg, Germany: Springer-Verlag; 1999. p. 164-179. DOI: 10.1.1.42.6458

[65] European Commission Directorate-General for Informatics. Data Quality Management. 2019. Available from: https://joinup.ec.europa.eu/sites/default/ files/document/2019-09/SEMIC [Accessed: July 9, 2021]

[66] Cichy C, Rass S. An overview of data quality frameworks. IEEE Access. 2019;**7**:24634-24648. DOI: 10.1109/ ACCESS.2019.2899751