MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-832

Harmonizing SNOMED CT with BioTopLite: An Exercise in Principled Ontology Alignment

Stefan Schulz[a], Catalina Martínez-Costa[a]

[a]*Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*

Abstract

The integration of heterogeneous ontologies is often hampered by different upper level categories and relations. We report on an on-going effort to align clinical terminology/ontology SNOMED CT with the formal upper-level ontology BioTop-Lite. This alignment introduces several constraints at the OWL-DL level. The mapping was done manually by analysing formal and textual definitions. Descriptive logic classifiers interactively checked mapping steps, using small modules for increasing performance. We present an effective workflow, using modules of several scales. However, only part of the classes and relations could easily be mapped. The implications for future evolution of SNOMED CT are discussed. It seems generally feasible to use a highly constrained upper-level ontology as an upper level for the benefit of future SNOMED CT versions that are more interoperable with other biomedical ontologies.

Keywords:

Biological Ontologies; SNOMED CT.

Introduction

The support of domain terminologies by formal ontologies is increasingly seen as an important requirement for semantic interoperability in health care and biomedical research. The integration between clinical and research data is essential for the advance of translational and personalized medicine. Both health care and biomedical research have terminology systems. Whereas the Unified Medical Language System (UMLS) [1] has focused on the integration of heterogeneous terminology systems in terms of lexical semantics, the discipline of Applied Ontology [2] has proposed to look beyond linguistic and conceptual structures. Applied Ontology's attention focuses on the referents, i.e., the entities denoted by terms and concepts, their ontological nature and the way they are related. Description logics (DLs) [3], especially the different OWL dialects [4] have become de-facto standards for ontologies.

An important aspect is the standardization of ontology artefacts in terms of upper-level ontologies**,** i.e., formal or semi-formal systems of categories, relations and axioms, linked to human-readable labels. The development of the two most well established upper-level ontologies, namely DOLCE [5] and BFO [6], focused on certain areas like cognitive sciences (DOLCE) and natural science (BFO). Other examples for upper-level ontologies in the biomedical domain include the GALEN upper level [7], the UMLS semantic network [8], and the OBO relation ontology RO [9].

The OBO Foundry effort [10] advocates the use of upper-level ontologies. In parallel, the ontological foundations of the large clinical terminology SNOMED CT [11] have substantially evolved, and the important field of medical classifications, with ICD as flagship, will be increasingly anchored in ontological grounds [12]. Whereas OBO Foundry ontologies are aligned with BFO 1.1 classes and with relations from RO, SNOMED CT's ontological framework is largely influenced by the legacy of former SNOMED versions, together with a frame-like concept model, which evolved almost untouched by formal-ontological deliberations.

An ontological upper-level shared by a range of different domain ontologies would have several advantages:

- Upper-level ontologies would force the categorization of domain entities into well-defined upper level categories. This is not at all trivial, regarding the inherent ambiguity of many legacy categories. A typical example is the pseudo-category *Problem*, a key concept in clinical documentation and decision-making, e.g., as single items that constitute clinical problem lists. *Problem* includes a broad range of statements on a person having a disease now or in the past, having a certain behaviour, having a history of surgical interventions etc., together with epistemic aspects like severity, significance, certainty, priority etc.
- Medical terms that denote what is commonly classified by *Disease* / *Disorder* turn out to be ambiguous [13]. For example, "allergy" may denote a disposition or a process, "fracture" may denote a damaged anatomical entity, but also the fracturing event, or even the life phase during which an organism exhibits the former and which has started with the latter.
- Upper-level ontologies would standardize the ways entities are related with each other. This standardization means a well-defined canon of relations. However, this situation requires a consensus about the precise meaning of relations, in terms of their algebraic properties and relation hierarchies, such as whether the relation **is part of** is reflexive, or whether **is part of** always implies **is included in**. The fact that important relations are ternary (e.g., *a* **is part of** *b* **at time** *t*), whereas the commonly used representation languages are binary is an issue not to be neglected [14]. This fact also complicates the current process of creation of an OWL version out of the upper level ontology BFO2 [15].
- Upper-level ontologies introduce constraining axioms that enforce conformance with domain ontologies that depend on them. Examples include domain / range constraints, such as that only members of the class *Process* can be in the domain of the relation **has participant**, but also further-reaching constraints such as that material entities can have material or immaterial entities as parts, whereas processed can only have processes as parts.

In this paper, we report on a first pilot of a harmonization effort that aims to fill the gap between SNOMED CT and other biomedical ontologies. We use the domain level ontology BioTopLite2 (BTL2) [16] in order to align SNOMED CT with well-defined classes and relations, and we use BTL2's rich set of axioms for validating the design decisions. Finally we will test the scalability of the map by measuring classification time, given that SNOMED CT mapped to BTL2 means to go from polynomial to exponential complexity.

Methods

SNOMED CT is gaining ground as an international clinical terminology standard. After the fusion of earlier SNOMED nomenclatures with the UK clinical terminology CTV3, the international standards organisation IHTSDO has embraced the mission of transforming SNOMED CT into a global healthcare language. Although the ontological SNOMED CT content, to date, is primarily released in relational format, a standard way of creating Description Logics (DL) axioms out of them has been described. This algorithm has been implemented in Perl script, from which the description logics axioms can be generated as an OWL-EL ontology [11]. The DL version has primarily been used to automatically generate taxonomic links in the SNOMED CT production process.

Table 1 shows the concepts of the uppermost level of the SNOMED CT concept hierarchy, under which all of the approximately 300,000 SNOMED CT concepts are grouped.

Table 1 – SNOMED CT upper concepts

Body structure	*Procedure*
Clinical finding	*Qualifier value*
Environment or geogr. location	*Record artefact*
Event	*Situation with explicit context*
Observable entity	*Social context*
Organism	*Special concept*
Pharmaceutical/biologic product	*Specimen*
Physical force	*Staging and scales*
Physical object	*Substance*

Table 2 shows the most frequent relations (object properties) in the OWL version (out of 62). These relations account for approximately 95% of all asserted relational statements in SNOMED CT axioms (totalling approx. 380,000).

Table 2 –most frequent SNOMED CT relations

Role group	**Direct device**
Finding site	**Direct substance**
Method	**Using substance**
Associated morphology	**After**
Procedure device	**Has specimen**
Has active ingredient	**Has focus**
Causative agent	**Finding context**
Has dose form	**Associated finding**
Interprets	**Has intent**
Procedure site - Indirect	**Procedure site**
Direct morphology	**Using access device**
Component	**Procedure context**
Has interpretation	**Associated procedure**
Occurrence	**Access**
Using device	**Specimen source topography**
Temporal context	**Laterality**
Subject relationship context	**Associated with**

BioTop was launched in 2006 as an Upper Domain Ontology in OWL DL. As BioTop never intended to compete with established ontologies, its developers created bridging ontologies to DOLCE, BFO, RO, and the UMLS Semantic Network and left the uppermost level deliberately flat. An important asset of BioTop has been its strong focus on constraining axioms, as an important mechanism for consistency checking, which is not yet available for BFO and RO. BioTop was used as a domain top level in several projects [16]. Later, a reduced form was released, named BioTopLite. Its classes and relations are shown in Tables 3 and 4.

Table 3 – BioTopLite2 upper classes

Disposition	*Process*
Function	*Quality*
Immaterial object	*Role*
Information object	*Temporal region*
Material object	*Value region*

Table 4 – BioTopLite2 upper relations (OWL object properties, without inverses, without descendants)

at some time	**includes**
causes	**precedes**
has condition	**projects onto**
has participant	**represents**

The **mapping process** was done for all SNOMED CT concepts and relations in Table 1 and 2. The use of the complete SNOMED CT OWL version for this exercise would have resulted in severe performance problems, slowing down the whole process.

We therefore decided to perform the mapping on modules of SNOMED CT. One type of modules (M1) was constructed based on signatures that contained one seed concept per SNOMED CT pattern. We defined a SNOMED pattern as a generalization of a subclass or equivalence axioms in the sense that each concept in the axiom was substituted by its uppermost ancestor, i.e., the concepts in Table 1. This method yielded 1,746 axiom types[1]. For each axiom type, a concept was selected randomly. The resulting set of concepts was used as a signature to create a module by following all outgoing links horizontally and vertically, according to [17]. Such a module has about 11,000 classes (variations resulting from random selection have only minor effects). Another series of modules (M2) was created from weighted (by subhierarchies) random signatures of different sizes.

The mapping was done completely manually, following previous work, employing an iterative approach [18,19]. The decision for a given map was done by thoroughly analysing the meaning of the candidate classes and relations, considering formal axioms as well as text definitions and hierarchical context. Each major ontology mapping step is checked by a DL reasoner, the results of which are then analysed and corrected under two perspectives: first, the classes tagged as 'inconsistent' are identified and the causes are investigated and repaired; second, whenever the ontology has reached a consistent state, the logical entailments are analysed for adequacy. Again, the causes are investigated and fixed, whenever wrong

1 Duplications of clauses like '**rel**$_m$ some *TopConcept*$_n$' were ignored; '**rel**$_m$ some *TopConcept*$_n$ and **rel**$_o$ some *TopConcept*$_p$' vs. '**rel**$_o$ some *TopConcept*$_p$ and **rel**$_m$ some *TopConcept*$_n$' and other variations in order were only counted once.

entailments were encountered. We used Protégé 5 [20] for editing, together with the HermiT [21] and FaCT++ reasoners [22], supported by the included explanation facility [23].

Results

The mapping of the classes yielded the following results (BioTopLite classes are indicated by the namespace prefix btl:, SNOMED CT by sct:)[2]:

- **Equivalence mapping** was possible only for the class sct:*Organism*.
- **Simple subclass mappings** were done for the following SNOMED CT top-level concepts: sct:*Event* under btl:*process*, sct:*Observable entity*, *Record artefact* and sct:*Staging and scales* under btl:*information object*, sct:*Pharmaceutical / biologic product*, sct:*Physical object* and sct:*Specimen* under btl:*polymolecular composite entity*, sct:*Physical force* under btl:*quality*, sct:*Procedure* under btl:*action*.
- **Complex subclass mappings**, i.e., those targeting compositional BTL2 expressions were done for sct:*Body structure*, which was mapped to the expression btl:*immaterial object* or btl:*structured biological entity*. Target for sct:*Clinical finding* was the expression btl:*disposition* or btl:*function* or btl:*material object* or btl:*process*. This expression was introduced for convenience as the defined class btl:*condition*, due to the necessity to pragmatically deal with the ambiguous, cross-category meaning of diseases, disorders and findings. sct:*Environment or geographic location* was mapped to btl:*immaterial object* or btl:*poly molecular composite entity*. Finally, sct:*Substance* was mapped to the disjunctive statement btl:*amount of pure substance* or btl:*compound of collective material entities*, due to the fact that also mixtures at molecular or microscopic level (blood, milk) are substances in SNOMED CT.
- **No mappings** resulted from the analysis of (i) sct:*Qualifier value*, (ii) sct:*Situation with explicit context*, (iii) sct:*Social context*, and (iv) sct:*Special concept*. In the first case, we found a large inhomogeneity, including actions like *Training – action*, qualities like *Decreased*, as well as numbers and units of measurement. The second case refers to entities of the frequently debated *Situation with explicit context* hierarchy, which is a kind of SNOMED-internal information model with concepts like *No temperature symptom* or *Treatment changed*. Although most of them could be mapped to btl:*Information object*, we refrained from a mapping. (For a thorough ontological analysis of these so-called context model concepts cf. [24]). Under (iii) we have encountered roles, individual humans, as well as population groups, and under (iv) we found inactive concepts and navigational concepts, which exist only to provide nodes in a navigation hierarchy, according to [25] and are therefore of no ontological relevance.

The mapping of relations proved far more complex, because of their larger number in both ontologies (see Tables 2 and 4) and the large number of constraining axioms attached to BTL2 relations. We decided to map not only the relations at the uppermost hierarchical level but also some of their descendants, because of their frequent use and special semantics. So far, we have limited the relation mapping effort to those relations that, together, cover 95% of all relational statements in SNOMED CT, see Table 2.

Only one relation equivalence was found, namely sct:**After** for btl:**precedes**.[3]

A special case was the sct:**RoleGroup** relation, which had originated as a mere syntactic construct to circumvent nested expressions [26], and then was found to be interpretable in several ways according to its context. We identified two different possible mappings: for domains of the type sct:*Clinical Finding* the corresponding BTL relation is btl:**has condition**, domains of the type sct:*Procedure* would be mapped to btl:**has part**.

Most SNOMED CT relations could be mapped as subrelations to BTL2 relations, together with refined domain and range restrictions. This is the case with the following relations:

- sct:**Finding site** mapped to btl:**is included in** with domain btl:*condition* and range sct:*Body structure;*
- sct:**Procedure site** mapped to btl:**is included in** with domain sct:*Procedure* and range sct:*Body structure;*
- sct:**Using device** mapped to btl:**has patient** with domain sct:*Procedure* and range sct:*Physical object*;
- sct:**Causative agent** mapped to btl:**caused by** with range sct:*Substance*;
- sct:**Laterality** mapped to btl:**bearer of** with range btl:*Quality*;
- sct:**Has active ingredient** mapped to btl:**has part** with domain sct:*Pharmaceutical / biologic product* and range sct:*Substance*;
- sct:**Associated morphology** mapped to btl:**has part** with domain btl:*condition* and range sct:*Morphologically abnormal structure*;
- sct:**Has dose form** mapped to btl:**is bearer of** with domain sct:*Pharmaceutical / biologic product* and range sct: *Drug dose form*;
- sct:**Has specimen** mapped to btl:**has patient** with domain sct:*Procedure* and range sct: *specimen*;
- sct:**Method** mapped to btl:**includes** with domain sct:*Procedure* and range sct:*Qualifier value*;
- sct:**Access** mapped to btl:**is bearer of** with domain range sct:*Qualifier value*;
- sct:**Occurrence** mapped to btl:**projects onto** with domain range sct:*Qualifier value*;
- sct:**Procedure device** mapped to btl:**has patient** with domain sct:*Procedure* and range sct:*Physical object*;
- sct:**Component** mapped to btl:**has patient** with domain sct:*Procedure* and range sct:*Substance*;
- sct:**Specimen source topography** mapped to the chain (btl:**at some time** • btl:**is included in**).

[2] We use *Italics* for classes / concepts and **Bold face** for relations (object properties)

[3] Another relation pair for which equivalence mapping seems plausible is sct:**part of** / btl:**is part of**. The former is not in our list as it is, to time, is infrequent SNOMED CT**.** After the planned redesign of the anatomy branch in SNOMED CT, this will be one of the most frequent relations.

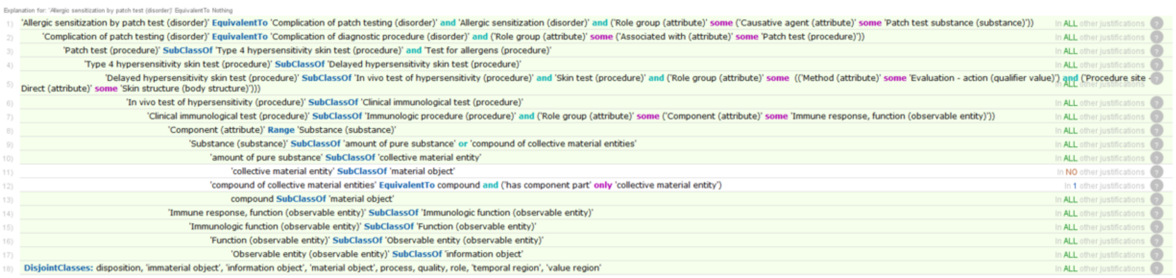

Figure 1 – Documentation of an unsatisfiable class using the Protégé explanation function

Several SNOMED CT subrelation introduce additional aspects, e.g. sct:**Procedure site - Indirect**, sct:**Direct morphology**, sct:**Direct device**, sct:**Using access device**. The differentiation of their meaning regarding their respective superrelations cannot be expressed by any BTL relation. One solution would be to assign roles or qualities to the target classes, e.g., the difference between direct or indirect surgical access.

There are complex relationships in SNOMED CT that cannot be mapped to any BTL2 relation. For instance sct:**Has focus** describes the intent of a procedure to reach a certain goal. An appropriate representation would require a model of intentionality, together with a reference to a class that is not (yet) instantiated. This would require an approach similar to the representation of plans, which is outside of what can be expressed in OWL-EL [27].

As argued in [24], the concepts in the SNOMED CT hierarchy *Situation in specific context* represent epistemic entities like components in an information model. As shown in [28], there are possibilities to express such statements in description logics, as instances of the class btl:*information object*, however requiring a more expressive logic than OWL-EL. We therefore exclude the following relations from our mapping: sct:**Finding context**, sct:**Procedure context**, sct:**Associated finding**, sct:**Interprets**, sct:**Has intent**, sct:**Subject relationship context**, sct:**Has interpretation**, sct:**Temporal context**, sct:**Subject relationship context**.

The mapping workflow turned out to critically depend on classification time. Whenever the ontology was in an inconsistent state, the classification of SNOMED CT modules of type M1 (in which all SNOMED CT patterns are represented) causes disruptions of more than fifteen minutes. In such cases, a switch to M2 modules of smaller scale was mostly sufficient for debugging, with classification times under two minutes. The debugging of insatisfiable classes was time consuming, which is illustrated by the output of the Protégé explanation function (Figure 1). At the end of the described phase, with a mapping of 95% of the relational clauses, there were on average three insatisfiable SNOMED CT concepts in modules with on average 11,050 concepts.

Conclusion

The on-going SNOMED CT – BTL2 aligning experience has shown up until now that large parts of SNOMED CT were compatible with a highly principled and compact upper level ontology like BioTopLite. The mapping at class level demonstrated a good agreement with BioTopLite for fifteen of the nineteen subhierarchy roots. That ontologically heterogeneous subhierarchies like sct:*Situation with specific context*, sct:*Qualifier value*, as well as sct:*Social context* and sct:*Special concept* could not be straightforwardly mapped was not surprising. We recommend that these SNOMED CT branches undergo a major redesign. At the level of the relations, the known ambiguity of the role group relation was confirmed. This construct requires further analysis and should eventually be substituted by a set of new relations that are clearly labelled. However, this would have to include nested relations into the SNOMED CT architecture, a desideratum that had been formulated on various occasions, by several stakeholders, such as the IHTSDO Observable group.

Regarding the other relations, SNOMED CT would be well served in most cases with a much smaller set of relations. The multitude of relations is, above all, a legacy issue. The simple BTL2 relation btl2:**is included in** would perfectly suit to substitute the SNOMED CT relations sct:**Finding Site** and sct:**Procedure Site**. The same would be true for btl:**has part**, which could substitute the relations sct:**Active ingredient** and sct:**Associated Morphology**. From an ontological point of view, we see no need to repeat the sort of the range concept (e.g., being a morphology or a substance) in the relation. This would simplify the construction and maintenance of SNOMED CT.

Although BTL2 uses the whole range of OWL-DL constructors, classification of SNOMED CT modules under BTL2 even of a size of 15,000 classes show satisfactory performance values, with about 15 minutes using the HermIT classifier. For a non-disruptive workflow it has been proven valuable to use small random modules that classify quickly after each modification of the map, because it is quite likely that the impact of a modelling error leads to insatisfiable classes even in these small modules. Once all small modules are satisfiable, then the more time consuming consistency check with a big module can be performed. Such modules that cover the complete variability of SNOMED CT by containing at least one class for each occurring design pattern have a size of approximately 11,000 classes. Checking the map with these modules might spot additional errors, or confirm the satisfiability documented by the smaller (incomplete) modules.

This approach could be integrated into the SNOMED CT maintenance and redesign workflow, which until now is mainly guided by the constraints formulated inside the frame-like SNOMED CT concept model. In contrast, the OWL version used in the production process cannot spot inconsistencies due to the inexpressiveness of the language.

In 2008, Rector and Brandt argued in favour of a more expressive description logics for SNOMED CT [29]. Our findings suggest the feasibility for this at least regarding the upper level.

Acknowledgements

This work has been funded by the SemanticHealthNet Network of Excellence within the EU 7th Framework Program, Call: FP7-ICT- 2011-7, agreement 288408. http://www.semantichealthnet.eu.

References

[1] United States National Library of Medicine. Unified Medical Language System (UMLS), 2015. http://www.nlm.nih.gov/research/umls.

[2] Munn K, Smith B, eds. Applied Ontology. An Introduction, Frankfurt am Main: Ontos, 2008.

[3] Baader F et al. The Description Logic Handbook (2nd Edition). Cambridge: Cambridge University Press, 2007.

[4] OWL2 profiles http://www.w3.org/TR/owl2-profiles.

[5] Borgo S, Masolo C. Ontological Foundations of DOLCE. Handbook on Ontologies (2nd Edition), Springer Verlag, 2009: 361-382.

[6] Grenon P, Smith B. (2004). SNAP and SPAN: Towards Dynamic Spatial Ontology. Spatial Cognition and Computation 4 (1): 69-103.

[7] Rector A, Rogers J. Patterns, Properties and Minimizing Commitment: Reconstruction of the GALEN Upper Ontology in OWL. Proc. of the EKAW*04 Workshop on Core Ontologies in Ontology Engineering, 2004, http://ceur-ws.org/Vol-118/.

[8] McCray AT. An upper-level ontology for the biomedical domain. Comp Funct Genomics. 2003; 4 (1): 80-84.

[9] Smith B et al. Relations in biomedical ontologies. Genome Biology 2005; 6 (5): R46. Epub 2005 Apr 28.

[10] Smith B et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 2007 Nov; 25 (11): 1251-1255.

[11] International Health Terminology Standards Development Organisation. SNOMED CT® Technical Implementation Guide. Copenhagen, Denmark, 2015. http://www.snomed.org/tig.pdf.

[12] Rodrigues JM et al. Sharing ontology between ICD 11 and SNOMED CT will enable seamless re-use and semantic interoperability. Stud Health Technol Inform. 2013;192:343-346.

[13] Schulz S et al. Scalable representations of diseases in biomedical ontologies. J Biomed Semantics, 2011, 17;2 Suppl 2:S6.

[14] Zamborlini V, Guizzardi G: On the Representation of Temporally Changing Information in OWL. EDOCW 2010: 283-292.

[15] Smith B et al. Basic Formal Ontology 2.0. Draft specification and user's guide. http://purl.obolibrary.org/obo/bfo/2012-07-20/Reference .

[16] Schulz S, Boeker M. BioTopLite: An Upper Level Ontology for the Life Sciences. Evolution, Design and Application. In: Furbach U, Staab S eds. Informatik 2013. IOS Press; 2013.

[17] Seidenberg J, Rector AL: Web ontology segmentation: analysis, classification and use. WWW 2006: 13-22.

[18] Schulz S, Hahn U. Medical knowledge reengineering – converting major portions of the UMLS into a terminological knowledge base. Int J Med Inform. 2001;64:207-221.

[19] Schulz S et al. Alignment of the UMLS semantic network with BioTop: methodology and assessment. Bioinformatics. 2009 Jun 15; 25 (12): i69-76.

[20] Stanford Center for Biomedical Informatics Research. The Protégé Ontology Editor and Knowledge Acquisition System, 2015. http://protege.stanford.edu.

[21] Shearer R, Motik B, Horrocks I, HermiT: a Highly-Efficient OWL Reason-er. 5th OWL-ED Workshop, 2008.

[22] Tsarkov D, Horrocks I. FaCT++ Description Logic Reasoner: System Description. In Third International Joint Conference on Automated Reasoning, 292-297. Springer, 2006.

[23] Horridge M, Parsia B, Sattler U. Laconic and precise justifications in OWL. Proc. of ISWC-08, volume 5318 of LNCS, 323-338, 2008.

[24] Martínez-Costa C, Schulz S. Ontology-based reinterpretation of the SNOMED CT context model. Proc. of the 4th International Conference on Biomedical Ontology, CEUR Workshop Proceedings, 2013; 1040: 90-95.

[26] Spackman KA, Dionne R, Mays E, Weis J. Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED. Proc AMIA Symp. 2002:712-716.

[27] Schulz S, Cornet R, Spackman K. Consolidating SNOMED CT's ontological commitment. Applied Ontology, 2011; 6: 1-11.

[28] Schulz S et al. An ontological analysis of reference in health record statements. FOIS 2014, 289-302.

[29] Rector AL, Brandt S. Why do it the hard way? The case for an expressive description logic for SNOMED. J Am Med Inform Assoc. 2008 Nov-Dec;15(6):744-751

Address for correspondence

Stefan Schulz
Institut für Medizinische Informatik,
Statistik und Dokumentation
Medizinische Universität Graz
Auenbruggerplatz 2/V
8020 Graz
Austria
stefan.schulz@medunigraz.at