The Journal of Open Source Software

snomedizer: R Interface to the SNOMED CT Terminology Server REST API

Peter Dutey-Magni[1] **and Anika Cawthorn**[1]

1 University College London

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

Electronic health record systems are expanding across the world and increasingly being used for operational planning and applied research. Electronic health records largely consist of unstructured data, which present challenges to analysts and researchers.

Solutions to these challenges lie in SNOMED CT, a global standard vocabulary for representing unstructured medical information. SNOMED CT terminology references and describes a wide range of concepts ranging from clinical anatomy to findings, procedures, and even medicines.

`snomedizer` is an R package to interrogate Snowstorm, the open-source SNOMED CT terminology server. It is designed for non-specialists and supports operations such as: extracting attributes of a target concept; reclassifying concepts into parent concepts; building codelists; and extracting basic information from free-text information. Providing access to these operations directly from a familiar data wrangling environment will lower the barrier/threshold to the SNOMED CT ontology faced by non-expert users.

Background

SNOMED CT is a clinical terminology system used in 80+ countries and present in over 70% of electronic medical records systems commercialised in Europe and North America (SNOMED International, 2021a, 2021b). Some jurisdictions now mandate its use in healthcare (New Zealand Ministry of Health, 2021; NHS Digital, 2016, 2021a, 2021b).

A extensive reader on SNOMED CT is available from Bhattacharyya (2016). Briefly, SNOMED CT includes a database of healthcare terms/synonyms in several languages. But it is also an ontology made up of 'concepts', i.e real-world entities defined in relation to each other by 'relationships' (see Figure 1). Relationships express a concept's inheritance (such as being a subtype of another concept) as well as its attributes (for example, a medical disorder can have a particular anatomical finding site).

SNOMED International has developed technical specifications for terminology services (SNOMED International, 2020), notably the Expression Query Language (ECL, SNOMED International, 2021c) which may be used to interrogate and retrieve information on concepts, their synonyms, and their relationships. Snowstorm (SNOMED International, 2021d) is the official implementation of these standards into a free and open-source Java/ElasticSearch application complete with a representational state transfer (REST) application programming interface.

The present paper sets out the design of `snomedizer`, an interface library providing access to the official SNOMED CT terminology service within the R programming environment.

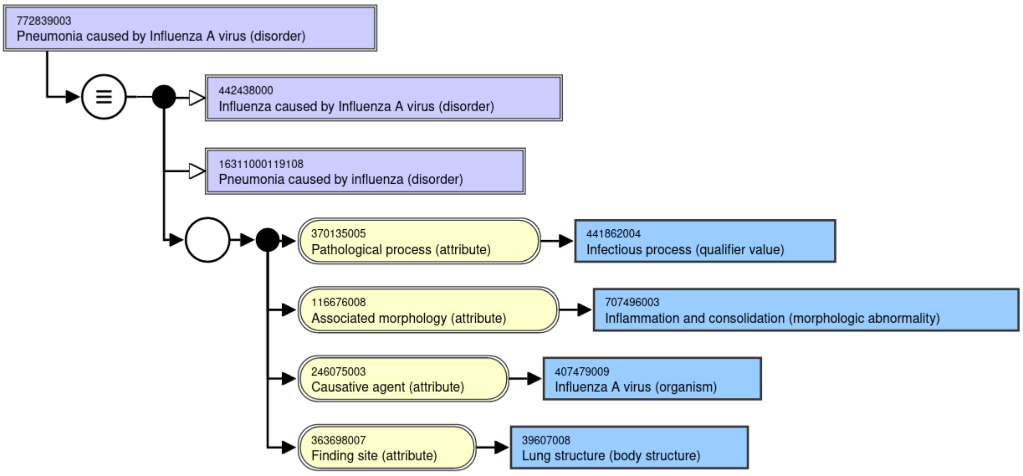
772839003
Pneumonia caused by Influenza A virus (disorder)
≡
442438000
Influenza caused by Influenza A virus (disorder)
16311000119108
Pneumonia caused by influenza (disorder)
370135005
Pathological process (attribute)
441862004
Infectious process (qualifier value)
116676008
Associated morphology (attribute)
707496003
Inflammation and consolidation (morphologic abnormality)
246075003
Causative agent (attribute)
407479009
Influenza A virus (organism)
363698007
Finding site (attribute)
39607008
Lung structure (body structure)

Figure 1: Diagram of stated relationships from SNOMED CT concept 772839003 | `Pneumonia caused by Influenza A virus (disorder)` | as produced by the SNOMED International browser

Statement of need

We considered a specific group of end users:

- healthcare analysts and health service researchers
- with data science skills foundations
- who need to use SNOMED CT in support of wider analytical and data wrangling tasks.

The adoption of SNOMED CT in healthcare analytics is hindered by three obstacles:

1. Building on-premise terminology services and loading/updating SNOMED CT releases is complex.
2. Substantial knowledge and skill are required to query a terminology server's application programming interface (API). Users must author HTTP requests and process their response. Typical operations may require multiple (sequential) calls to the API, requiring additional programming.
3. Terminology services do not offer many hands-on training resources with real-life examples.

Typical tasks (use cases) performed by the target end users include:

- building codelists (Davé & Petersen, 2009; Watson, Nicholson, Hamilton, & Price, 2017)
- rapidly reclassifying concepts into higher-level categories (ascendants), eg `312371005 | Acute infective bronchitis (disorder) |` to `50417007 | Lower respiratory tract infection (disorder) |`
- extracting characteristics of a concept, for instance the `363698007 | Finding site (attribute) |` of a disorder, or the `127489000 | Has active ingredient (attribute) |` of a medical product.

Design of snomedizer

Snowstorm (SNOMED International, 2021d) already largely addresses obstacle (1) above. It enjoys a wide user base, and serves as the terminology service for many other SNOMED services (such as the SNOMED International browser). This provides welcome community support, for example when facing difficulties loading particular terminology extensions.

`snomedizer` (Dutey-Magni & Cawthorn, 2021) is intended to address obstacles (2) and (3) by providing easy functions to send and retrieve common queries to and from an existing Snowstorm terminology service.

Key requirements of `snomedizer` are:

- The software must not require advanced knowledge of ontological reasoning, natural language processing or software engineering.
- The software must be free and open-source.
- The software must be interoperable with popular data wrangling software used by the target user group. R is a leading leading health data science language along with Python (Meyer, 2019), with a strong community (https://nhsrcommunity.com/). R use has been popularised by the data wrangling library `dplyr` and tidy data design principles (Wickham, 2014).
- The software must help retrieve ascendants/descendants of one or more concepts.
- The software must help retrieve attributes of one or more concepts.
- The software must support bespoke and complex queries devised by the user.
- The software must provide extensive documentation.
- Users must be able to practice using tutorials and examples without building a Snowstorm endpoint.
- The software must warn users of potential incompatibility with the Snowstorm endpoint they choose to query.

To fulfill these requirements, `snomedizer` provides six wrapper functions providing user-friendly access to common operations:

- `concept_ancestors()` and `concept_descendants()` fetch active ancestors/descendants of one or more concepts
- `concept_descriptions()` fetches descriptions of one or more concepts
- `concept_find()` searches SNOMED CT concepts by term, ECL query, or concept identifiers
- `concept_included_in()` determines whether one or more concept are subtypes of a target set of concepts
- `concept_map()` maps SNOMED CT concepts to other terminology or code systems (using map reference sets).

To conform with tidy data principles (Wickham, 2014), these wrapper functions return results as data frames and support vectors inputs. They also specify default options relevant to most users and trigger warnings/errors. `snomedizer` also includes 22 functions (prefixed with `api_*()`) providing direct implementations of relevant Snowstorm API operations (terminology authoring and maintenance operations are not supported), and a range of utility functions for handling connections to Snowstorm endpoints.

A companion website contains all documentation and tutorials (https://snomedizer.web.app/). Tutorials make use of public Snowstorm endpoints maintained by SNOMED International. These endpoints can be discovered with the `snomed_public_endpoint_suggest()` function. Their use is for reference only and subject to the SNOMED International Browser License Agreement (https://browser.ihtsdotools.org/).

HTTP requests to Snowstorm have the potential to disclose confidential personal data, for instance when containing free-text terms or rare and sensitive concepts (eg human immunodeficiency virus infections). Organisations looking to process personal information can build Snowstorm and use `snomedizer` on premise behind a firewall to preserve confidentiality.

Release 0.3.0 of `snomedizer` supports Snowstorm versions 7.6.0-7.9.3. Compatibility with future releases of Snowstorm is be monitored through a review of release notes and a range of regression tests. `snomedizer` releases are documented in a change log and state the latest Snowstorm release they are known to support. The package generates user warnings if the configured endpoint is not fully compatible with the current version of `snomedizer`. Bug reports and requests for new features are encouraged and can be submitted on the package repository (https://github.com/ramses-antibiotics/snomedizer/issues/).

Acknowledgements

This project was funded by the National Institute for Health Research (NIHR) [Antimicrobial Resistance Capital Funding award 200652]. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

References

Bhattacharyya, S. B. (2016). *Introduction to SNOMED CT*. Springer. Retrieved from https://doi.org/10.1007/978-981-287-895-3_7

Davé, S., & Petersen, I. (2009). Creating medical and drug code lists to identify cases in primary care databases. *Pharmacoepidemiology and Drug Safety*, *18*(8), 704–707. doi:https://doi.org/10.1002/pds.1770

Dutey-Magni, P., & Cawthorn, A. (2021). *snomedizer: R Interface to the SNOMED CT Terminology Server REST API*. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.5705568

Meyer, M. A. (2019). Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings. *Journal of the American Medical Informatics Association*, *26*(5), 383–391. doi:10.1093/jamia/ocy181

New Zealand Ministry of Health. (2021). *HISO 10048 emergency care data standard*. Retrieved from https://www.health.govt.nz/publication/hiso-10048-emergency-care-data-standard

NHS Digital. (2016). *SCCI0034 SNOMED CT Requirements Specifications Amd 35/2016*. Retrieved from https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/scci0034-snomed-ct

NHS Digital. (2021a). *DAPB4013: Medicine and Allergy/Intolerance Data Transfer Requirements Specification (Amd 5/2021)*. Retrieved from https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-includin publications-and-notifications/standards-and-collections/dapb4013-medicine-and-allergy-intoleran

NHS Digital. (2021b). *DAPB4017: Pathology Test and Results Standard Specification (Amd 44/2020)*. Retrieved from https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/dapb4017-pathology-test-and-results-sta

SNOMED International. (2020). *SNOMED CT Terminology Services Guide. Publication date: 2020-09-30.* Retrieved from http://snomed.org/tsg

SNOMED International. (2021a). *2020 Annual Report.* Retrieved from https://www.paperturn-view.com/?pid=MTY165474

SNOMED International. (2021b). *SNOMED CT: Articulating Stakeholder Value.* Retrieved from https://www.paperturn-view.com/?pid=MTU155774

SNOMED International. (2021c). *SNOMED CT Expression Constraint Language Specification and Guide. Version 1.6. Publication date: 2021-10-05.* Retrieved from http://snomed.org/ecl

SNOMED International. (2021d). *Snowstorm: SNOMED CT Terminology Server Using Elasticsearch.* Retrieved from https://github.com/IHTSDO/snowstorm

Watson, J., Nicholson, B. D., Hamilton, W., & Price, S. (2017). Identifying clinical features in primary care electronic health record studies: Methods for codelist development. *BMJ Open*, *7*(11). doi:10.1136/bmjopen-2017-019637

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23. doi:10.18637/jss.v059.i10