Redundant Elements in SNOMED CT Concept Definitions

Kathrin Dentler[1,2,⋆] and Ronald Cornet[2,3,⋆⋆]

[1] Dept. of Computer Science, VU University Amsterdam, The Netherlands
[2] Dept. of Medical Informatics, Academic Medical Center,
University of Amsterdam, The Netherlands
k.dentler@vu.nl
[3] Department of Biomedical Engineering, Linköping University, Sweden

Abstract. While redundant elements in SNOMED CT concept definitions are harmless from a logical point of view, they unnecessarily make concept definitions of typically large ontologies such as SNOMED CT hard to construct and to maintain. In this paper, we apply a fully automated method to detect intra-axiom redundancies in SNOMED CT. We systematically analyse the completeness and soundness of the results of our method by examining the identified redundant elements. In absence of a gold standard, we check whether our method identifies concepts that are likely to contain redundant elements because they become equivalent to their stated subsumer when they are replaced by a fully defined concept with the same definition. To evaluate soundness, we remove all identified redundancies, and test whether the logical closure is preserved by comparing the concept hierarchy to the one of the official SNOMED CT distribution. We found that 35,010 of the 296,433 SNOMED CT concepts (12%) contain redundant elements in their definitions, and that the results of our method are sound and complete with respect to our partial evaluation. We recommend to free the stated form from these redundancies. In future, knowledge modellers should be supported by being pointed to newly introduced redundancies.

Keywords: SNOMED CT, OWL 2 EL, Redundancies, Reasoning.

1 Introduction

SNOMED Clinical Terms (SNOMED CT) allows for meaning-based recording and retrieval of clinical information, which thereby becomes (re)usable. One of the advantages of SNOMED CT is its large size and coverage, which on the other hand makes defining new and maintaining existing concepts a challenging task.

⋆ corresponding author.

⋆⋆ Remark: Ronald Cornet is a member of the Technical Committee of the International Health Terminology Standards Development Organization (IHTSDO), which publishes SNOMED CT. His position at the IHTSDO, however, had no bearing on the research study or results.

N. Peek, R. Marín Morales, and M. Peleg (Eds.): AIME 2013, LNAI 7885, pp. 186–195, 2013.

Various (automated) auditing methods have been developed that can be applied to the content of controlled biomedical terminologies, amongst others to ensure the quality factor non-redundancy [16]. While such methods mostly aim at detecting equivalent concepts, also parts or elements of concept definitions, i.e. intra-axiom redundancies, are problematic. The detection of intra-axiom redundancies is required during design time. In fact, Spackman et al. reported back in 2001 that "... during the concept definition process there has been confusion among modelers about which roles need to be explicitly modeled and which ones can be left unstated. Some of this confusion arises because of uncertainty about which roles and values are inherited from supertypes" [13]. And even though redundancies are harmless from a logical point of view, they impede the maintainability of a terminology [11], [8], as they misleadingly suggest that new information has been added to a concept, while in reality, this "new" information is more general than or equivalent to information that already has been stated in the definition of the same concept or a superconcept. In this paper, we make an inventory of redundant elements in SNOMED CT concept definitions.

2 Background

2.1 SNOMED CT Concept Definitions and Rolegroups

SNOMED CT is based on the lightweight Description Logic EL^+ [1]. Its concepts are defined by conjunctions of other concepts as well as role-value pairs which are represented as exists restrictions ($\exists$), and can be either ungrouped or grouped in so-called *rolegroups* [5]. In SNOMED CT, rolegroups allow to nest or rather group existential restrictions within an existential restriction on a role named rolegroup. Concepts can be either *primitive*, i.e. specified by *necessary* conditions only (denoted by the subsumption operator $\sqsubseteq$) or *fully defined*, i.e. specified by both *necessary and sufficient* conditions (denoted by the equivalence operator $\equiv$). Example 1 presents a fully defined sample concept, which is defined by the conjunction of one concept and two rolegroups.

Example 1 (Brain stem contusion with open intracranial wound. RG stands for rolegroup).

```
Brain stem contusion with open intracranial wound ≡
    Contusion of brain with open intracranial wound ⊓
    ∃RG(∃Associated morphology.Open wound ⊓
        ∃Finding site.Intracranial structure) ⊓
    ∃RG(∃Associated morphology.Open contusion ⊓
        ∃Finding site.Brainstem structure)
```

2.2 Trivial and Non-trivial Primitive Concepts

For our evaluation, we distinguish *trivial primitive concepts*, that are primitive and subsumed by one concept only, and *non-trivial primitive concepts*, that are primitive and described by the conjunction of several concepts and optional additional exists restrictions. With regard to Example 2, we refer to the concept

Brain tissue structure as trivial primitive, and to *Structure of lobe of brain* as non-trivial primitive.

Example 2 (Structure of lobe of brain).

```
Brain tissue structure ⊑ Brain part

Structure of lobe of brain ⊑
    Brain part ⊓ Brain tissue structure
```

2.3 Redundant Elements in SNOMED CT Concept Definitions

An element that is part of a concept definition, i.e. a concept or an existential restriction, is redundant if it has been stated explicitly even though it is already implied by the definition of the same concept or a stated superconcept. Therefore, we define an element to be redundant if it is more general than or equivalent to an element that is contained in the definition of the same concept or a stated superconcept. Redundant elements can be eliminated without affecting the ontology's logical closure. For example, the concept *Brain part* in the definition of the concept *Structure of lobe of brain* in Example 2 is redundant as it subsumes the concept *Brain tissue structure*.

3 Materials and Methods

We employed the July 2012 version of SNOMED CT in Release Format 2, which was transformed to OWL with the Perl script released in the same version. The script makes use of the released concept and stated relationships tables. The latter represents the faithful representation of the information entered by modellers.

We relied on the high-performance reasoner ELK [15] to classify SNOMED CT, and to check for subsumption and equivalence relationships between concepts and roles, while Pellet [12] was used in our evaluation to explain equivalence relationships that were hard to reproduce manually. We relied on the OWL API [9] to carry out all experiments.

3.1 Method to Detect Redundant Elements in SNOMED CT Concept Definitions

We exploit the simple structure of SNOMED CT and its rolegroups to detect intra-axiom redundancies. Therefore, we adapted and extended the rules 1 to 3 of redundancy elimination for concept definitions that contain rolegroups as defined by Spackman et al. [14] (and adopted their original numbering). The rules are based on Definition 1.

Definition 1. More general or equivalent exists restriction. *An exists restriction is more general than or equivalent to another exists restriction whenever both its role and its value concept subsume or are equivalent to the respective elements in the other exists restriction.*

$\exists R.C \sqsupseteq \exists S.D \iff (R \sqsupseteq S)$ and $(C \sqsupseteq D)$

All concept definitions are merely conjunctions of ungrouped or grouped exists restrictions and superconcepts. Therefore, the rules define for each of these elements whether they are redundant:

1. An ungrouped exists restriction is redundant when it is more general than or equivalent to an ungrouped exists restriction within the definition of *the same concept or a superconcept.*
   $(\exists \mathtt{R.C} \sqcap \exists \mathtt{S.D} \sqcap \exists \mathtt{T.E}) \equiv (\exists \mathtt{S.D} \sqcap \exists \mathtt{T.E}) \Longleftrightarrow \exists \mathtt{R.C} \sqsupseteq \exists \mathtt{S.D}$

2. A rolegroup is redundant when all its exists restrictions are more general than or equivalent to those contained in another rolegroup in the definition of *the same concept or a superconcept.*
   $(\mathtt{RG}(\exists \mathtt{R}_1\mathtt{.C}_1 \sqcap \ldots \sqcap \exists \mathtt{R}_n\mathtt{.C}_n) \sqcap \mathtt{RG}(\exists \mathtt{S}_1\mathtt{.D}_1 \sqcap \ldots \sqcap \exists \mathtt{S}_m\mathtt{.D}_m)) \equiv \mathtt{RG}(\exists \mathtt{S}_1\mathtt{.D}_1 \sqcap \ldots \sqcap \exists \mathtt{S}_m\mathtt{.D}_m)$
   $\Longleftrightarrow \forall \mathtt{i=1,\ldots,n}\ \exists \mathtt{j=1,\ldots,m} \mid \exists \mathtt{R}_i\mathtt{.C}_i \sqsupseteq \exists \mathtt{S}_j\mathtt{.D}_j$

3. An exists restriction is redundant within a rolegroup when it is more general than or equivalent to another exists restriction in *the same rolegroup.*
   $\mathtt{RG}(\exists \mathtt{R.C} \sqcap \exists \mathtt{S.D} \sqcap \exists \mathtt{T.E}) \equiv \mathtt{RG}(\exists \mathtt{S.D} \sqcap \exists \mathtt{T.E}) \Longleftrightarrow \exists \mathtt{R.C} \sqsupseteq \exists \mathtt{S.D}$

4. A concept is redundant when it is more general than or equivalent to one of the other concepts in the definition of *the same concept or a superconcept.*
   $(\mathtt{C} \sqcap \mathtt{D}) \equiv \mathtt{D} \Longleftrightarrow \mathtt{C} \sqsupseteq \mathtt{D}$

Rule 3 is an exception with regard to our redundancy definition, as it does not concern an element of a concept definition, but an element within an element. To test whether a concept is defined redundantly, these four rules are applied to a concept and all its stated superconcepts. As the rules are independent from each other, their execution order should not influence the obtained results.

3.2 Evaluation of Our Method

To evaluate the results obtained by the application of the four rules of redundancy detection, we assess the completeness and soundness of its output. In absence of a gold standard, we measure completeness by matching our findings to definitions that are likely to be redundant according to Cornet's and Abu-Hanna's method [4], and soundness by checking whether the logical closure is preserved after classifying the manipulated version of the ontology.

Completeness: Comparison of Identified Redundant Concepts to Redundant Concepts According to Cornet's and Abu-Hanna's Method. Cornet's and Abu-Hanna's method [4] detects concepts with equivalent definitions in terminological systems represented in a description logic, to addresses the problems of redundancy and underspecification. Concepts that become equivalent to any superconcept when applying this method are likely to be defined redundantly [3]. Let us regard Example 3, which presents a sample group of equivalent concepts that can be detected by applying this method.

Example 3 (Group of concepts with equivalent concept definitions).

```
Finding of volume of heart sounds ⊑
    Finding of heart sounds ⊓
        ∃RG(∃Interprets.Loudness of heart sounds)

Heart sounds diminished ⊑
    Finding of volume of heart sounds ⊓
        ∃RG(∃Finding site.Heart structure)

Heart sound volume variable ⊑
    Finding of volume of heart sounds ⊓
        ∃RG(∃Finding site.Heart structure)

Heart sound inaudible ⊑
    Finding of volume of heart sounds ⊓
        ∃RG(∃Finding site.Heart structure)
```

Here, we can make two interesting observations. First, we see three concepts with definitions that obviously become equivalent when making these concepts fully defined. Second, the three concepts become equivalent to their superconcept *Finding of volume of heart sounds*, and thus, they are likely to be defined redundantly. And indeed, four steps up the concept hierarchy, we encounter their common superconcept presented in Example 4, which already contains a rolegroup that defines the *Finding site* to be the *Heart structure*.

Example 4 (Explanation for redundancy).

```
Cardiac finding ⊑
    Cardiovascular finding ⊓
        ∃RG(∃Finding site.Heart structure)
```

We evaluate the results obtained by the application of the four rules of redundancy detection by checking whether the concepts that are likely to be redundant according to Cornet and Abu-Hanna are indeed contained in the identified set of redundant concepts. In order to detect redundant definitions, we apply the approach proposed by Cornet and Abu-Hanna as follows:

1. Replace each non-trivial primitive concept by a fully defined concept with the same definition.
2. Classify the ontology.
3. For each concept in the ontology, retrieve equivalent concepts from reasoner.
4. Identify concepts that have become equivalent to any stated superconcept, as those are likely to be defined redundantly.
5. Identify and exclude indirect redundancies that emerge due to concepts being subsumed by the conjunction of concepts with equivalent definitions such as in Example 5 and wrongly identified redundancies due to the propagation of equivalence such as in Example 6.[1]

[1] Please note that these cases could be prevented by applying the method only on one superconcept - subconcept pair at a time instead of the entire SNOMED CT. We did not apply this method because it is not feasible even with very fast classification times.

Example 5 (Concepts without intra-axiom redundancy: Because Midwifery personnel *and* Professional midwife *have the same definitions, they become equivalent. And because* Auxiliary midwife *is being subsumed by the two of them, it also becomes equivalent.).*

```
Auxiliary midwife ⊑
    Professional midwife ⊓ Midwifery personnel

Professional midwife ⊑
    Medical, dental, veterinary/related worker ⊓
    Health visitor, nurse/midwife

Midwifery personnel ⊑
    Medical, dental, veterinary/related worker ⊓
    Health visitor, nurse/midwife
```

Please note that Cornet's and Abu-Hanna's method does not necessarily retrieve all redundant concepts. For example, a concept can refine its stated superconcept and additionally contain redundant elements. Likewise, redundant elements in fully defined concept definitions are not detected by Cornet's and Abu-Hanna's method. Therefore, the evaluation of the results of the four rules of redundancy detection can only be partial.

Example 6 (Example for wrongly identified redundancy. The concepts Pancreatic function outside reference range *and* Measurement finding outside reference range *would be equivalent if all involved concepts were fully defined.).*

```
Pancreatic function outside reference range ⊑
    Measurement finding outside reference range ⊓
    ∃RG(∃Has interpretation.Outside reference range ⊓
    ∃Interprets.Pancreatic function test)

Measurement finding outside reference range ≡
    Measurement finding ⊓
    ∃RG(∃Has interpretation.Outside reference range ⊓
    ∃Interprets.Measurement procedure)

Pancreatic function test ⊑
    Measurement procedure ⊓
    ∃RG(∃Has Method.Measurement - action)

Measurement procedure ≡
    Procedure by method ⊓
    ∃RG(∃Has Method.Measurement - action)
```

Soundness: Preservation of Logical Closure. Deleting redundant parts of concept definitions should not affect the logical closure, and therefore a change in the concept hierarchy would indicate the removal of a non-redundant part of a concept definition. Thus, we delete all identified intra-axiom redundancies and check whether the computed concept hierarchy obtained from classifying the manipulated version is the same as the one obtained from classifying the original version by bi-directional comparison of both versions to the official SNOMED CT distribution.

4 Results: Redundant Elements in Concept Definitions

Applying the four rules of redundancy detection, 35,010 of the 296,433 SNOMED CT concepts (12%) were identified to contain redundant elements in their definitions. Table 1 gives an overview of the results, only regarding the first explanation for these redundancies (the rules were applied in the same order as they are presented in this paper). 11,858 of these concepts are fully defined, and 23,152 non-trivial primitive.

Example 7 (Parenteral form thymoxamine).

```
Parenteral form thymoxamine (product) ≡
    Thymoxamine (product) ⊓
        ∃Has active ingredient.Thymoxamine (substance)

Thymoxamine (product) ⊑
    Alpha blocking vasodilator ⊓ Alpha 1 adrenergic blocking agent ⊓
        ∃Has active ingredient.Thymoxamine (substance)
```

Table 1. Detected concepts with redundant elements. The examples in column 'example' refer to the examples disseminated along the paper.

Rule	Concepts	Example and Explanation
1 (ungrouped exists restriction)	7,874	Example 7: The ungrouped exists restriction $\exists$*Has active ingredient.Thymoxamine (substance)* is redundant, as it is already contained in the superconcept *Thymoxamine (product)*.
2 (rolegroup)	26,599	Example 1: The first rolegroup is redundant, as it is more general than the second one, because *open wound* subsumes *open contusion*, and *Intracranial structure* subsumes *Brainstem structure*.
3 (grouped exists restriction)	6	Example 8: The exists restriction $\exists$*Associated morphology.Traumatic abnormality* in the first rolegroup is redundant, as *Traumatic abnormality* subsumes *Closed traumatic abnormality*.
4 (concept)	531	Example 2: The concept *Brain part* is redundant as it subsumes the concept *Brain tissue structure*.

Example 8 (Closed skull fracture with intracranial injury).

```
Closed skull fracture with intracranial injury ≡
    Fracture of skull ⊓
        ∃RG(∃Finding site.Intracranial structure ⊓
        ∃Associated morphology.Traumatic abnormality ⊓
        ∃Associated morphology.Closed traumatic abnormality) ⊓
        ∃RG(∃Associated morphology.Fracture, closed ⊓
        ∃Finding site.Bone structure of cranium)

Explanation:
Closed traumatic abnormality ⊑ Traumatic abnormality
```

Figure 1 shows the SNOMED CT categories that the concepts with redundant elements belong to. Figure 2 depicts the distances between redundant concepts and the concepts containing the explanation for the redundancy. A distance of 0 is interesting as it makes a concept redundant with regard to its own definition. But also long distances are interesting: an element is introduced, not repeated

for some concepts down the hierarchy, but then it is. The concept *Measurement of Human T-lymphotropic virus 1 recombinant glycoprotein 21 antibody and Human T-lymphotropic virus 2 recombinant glycoprotein 21 antibody* is among the concepts with the longest distance to its explanation (9 steps).

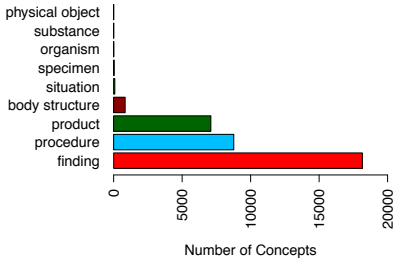

Fig. 1. SNOMED CT categories of concepts with redundancies

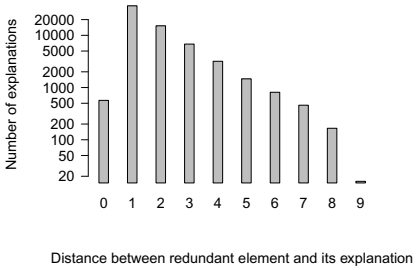

Fig. 2. Distances between redundant concepts and the concepts containing the explanation

An exhaustive search for all redundant elements and all explanations results in 65,336 explanations: 13,808 for rule 1, 50,680 for rule 2, 6 for rule 3 and 842 for rule 4. The maximum number of explanations is 16 for the concept *Late congenital syphilitic meningitis*. The concept with the most (6) redundant elements is *Diphtheria + tetanus + pertussis + poliomyelitis + recombinant hepatitis B virus + recombinant haemophilus influenzae type B vaccine*.

5 Evaluation

5.1 Completeness

Applying Cornet's and Abu-Hanna's method, 45,975 concept definitions with at least one other concept with a logically equivalent definition have been identified, containing a total of 12,823 non-trivial primitive concepts with definitions that are equivalent to the definition of at least one of their stated superconcepts.

12,094 of these redundancies have been confirmed to be redundant by our method to detect intra-axiom redundancies. 698 out of the 729 non-confirmed redundancies were subsumed by the conjunction of concepts, such as the concept in Example 5. For the remaining 31 non-confirmed redundancies, we successfully generated explanations with Pellet based on the manipulated version of SNOMED CT. A manual revision confirmed that all of the explanations contained further axioms that have been re-defined from being primitive to fully defined, such as the explanation given in Example 6. Therefore, the results of our method are complete with regard to Cornet's and Abu-Hanna's method.

5.2 Soundness

We generated the logical closure of both the original and the manipulated OWL versions of SNOMED CT, and compared the computed class hierarchies to the one contained in the official distribution. The OWL versions and the database table contained exactly the same set of 438,554 subclass axioms or respectively "is-a" relations.

6 Related and Future Work

In the past, most proposed methods focused at the detection of truly redundant, i.e. equivalent, concepts. Cimino has developed a method to identify multiple synonymous concepts and applied it to the 2001 UMLS Metathesaurus [2]. Grimm and Wissmann [8] provide methods to compute irredundant ontologies, and Entendre [6] makes users aware of redundancies.

The IHTSDO[2] describes methods to convert concepts into normal forms, some of which imply the elimination of redundancies, and Peng et al. [10] have proposed a method to identify redundant classifications, i.e. unnecessary, simultaneous assignments to sub- and superconcepts. The Ecco tool [7] facilitates the analysis of ontology differences by applying methods to syntactically or semantically detect effectual changes as well as ineffectual changes such as adding or deleting intra-axiom redundancies.

An interesting direction of future work would be to generalise our method. In principle, our definition of a redundant element could be operationalised directly by checking whether an element is more general than or equivalent to an element that is contained in the definition of the same concept or a stated superconcept.

7 Discussion and Conclusions

Our results show that 35,010 of all 296,433 SNOMED CT concepts (12%) are defined redundantly. These redundancies unnecessarily impede the work of knowledge modellers, and our own experience confirms that manual search for the causes of redundancies can be a tedious task. Therefore, we suggest to remove them from the stated relationships. To reach this goal, the four rules of redundancy detection would have to be applied to the entire SNOMED CT once.[3] Further redundancies should be avoided by pointing knowledge modellers to newly introduced redundancies in the definitions of the concepts they are currently working on, and explaining why these elements are redundant. As shown by Figure 2, most redundant elements are so due to nearby superconcepts, so that the explanations will most probably be intuitive. For this task, the four

[2] `http://www.ihtsdo.org/`

[3] It should be noted that applying the four rules of redundancy detection to the entire SNOMED CT is computationally expensive (ca. 6 hours on a laptop equipped with a 2.8 GHz Intel Core 2 Duo processor and 8 GB of physical memory). However, analysing only one concept is sufficiently fast to be executed as a background process.

rules of redundancy detection could be applied as a background process of terminology editing tools to the concepts that are currently being edited. In order to support these goals, we make both our tools and our results freely available[4].

References

1. Baader, F., Lutz, C., Suntisrivaraporn, B.: Is tractable reasoning in extensions of the description logic EL useful in practice? In: Proceedings of the 2005 International Workshop on Methods for Modalities, pp. 1–26 (2005)
2. Cimino, J.J.: Battling Scylla and Charybdis: the search for redundancy and ambiguity in the 2001 UMLS metathesaurus. In: Proceedings of the AMIA Symposium, pp. 120–124 (January 2001)
3. Cornet, R., Abu-Hanna, A.: Two DL-based methods for auditing medical terminological systems. In: AMIA Symposium, pp. 166–170 (January 2005)
4. Cornet, R., Abu-Hanna, A.: Auditing description-logic-based medical terminological systems by detecting equivalent concept definitions. International Journal of Medical Informatics 77(5), 336–345 (2008)
5. Cornet, R., Schulz, S.: Relationship groups in SNOMED CT. Stud. Health Technol. Inform. 31(0), 223–227 (2009)
6. Denaux, R., Thakker, D., Dimitrova, V., Cohn, A.: Entendre: Interactive Semantic Feedback for Ontology Authoring, `files.ifi.uzh.ch`
7. Gonçalves, R., Parsia, B., Sattler, U.: Ecco: A Hybrid Diff Tool for OWL 2 ontologies. In: Proceedings of the 9th International Workshop on OWL: Experiences and Directions (2012)
8. Grimm, S., Wissmann, J.: Elimination of redundancy in ontologies. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part I. LNCS, vol. 6643, pp. 260–274. Springer, Heidelberg (2011)
9. Horridge, M., Bechhofer, S.: The OWL API: A Java API for Working with OWL 2 Ontologies. In: 6th OWL Experienced and Directions Workshop (2009)
10. Peng, Y., Halper, M.H., Perl, Y., Geller, J.: Auditing the UMLS for redundant classifications. In: AMIA Symposium, pp. 612–616 (January 2002)
11. Schlobach, S., Cornet, R.: Logical support for terminological modeling. Studies in Health Technology and Informatics 107, 439–443 (2004)
12. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical OWL-DL reasoner. Web Semantics: Science, Services and Agents on the World Wide Web 5(2), 51–53 (2007)
13. Spackman, K.A.: Normal forms for description logic expressions of clinical concepts in SNOMED RT. In: Proceedings of the AMIA Symposium, pp. 627–631 (January 2001)
14. Spackman, K.A., Dionne, R., Mays, E., Weis, J.: Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED. In: Proceedings of the AMIA Symposium, 712–746 (January 2002)
15. Kazakov, Y., Krötzsch, M., Simančík, F.: Concurrent Classification of $\mathcal{EL}$ Ontologies. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 305–320. Springer, Heidelberg (2011)
16. Zhu, X., Fan, J.-W., Baorto, D.M., Weng, C., Cimino, J.J.: A review of auditing methods applied to the content of controlled biomedical terminologies. Journal of biomedical informatics 42(3), 413–425 (2009)

[4] `https://github.com/kathrinrin/redundancies`