Efficient construction of a new ontology for life sciences by sub-classifying related terms in the Japan Science and Technology Agency thesaurus

Tatsuya Kushida[1*,] Kouji Kozaki[2] Yuka Tateisi[1] Katsutaro Watanabe[3] Takeshi Masuda[2] Katsuji Matsumura[1,3] Takahiro Kawamura[3] and Toshihisa Takagi[1,4]

[1] National Bioscience Database Center, Japan Science and Technology Agency, Tokyo, Japan
[2] The Institute of Scientific and Industrial Research, Osaka Univ., Ibaraki, Japan
[3] Dept. of Information Planning, Japan Science and Technology Agency, Tokyo, Japan
[4] Dept. Biological Sciences, Grad. School of Science, The Univ. of Tokyo, Tokyo, Japan

ABSTRACT

We are developing a new ontology for life sciences that can be used to interlink biological concepts from various categories with approximately 10,000 concepts and 31 types of relations. We create these relations by sub-classifying the related terms (RT) that are used in the thesaurus of Japan Science and Technology Agency (JST) for associating concepts along with the broader and narrower terms. In this study, we describe an efficient ontological development method based on the JST thesaurus in terms of the majority decision of a panel of life-sciences experts. Three trained curators sub-classified 2850 RTs into 31 types of relations an improved version of Hozo ontology editor. We evaluated the results and confirmed high precision (0.93) and recall (0.83). Finally, a manager adjudicated the results by the curators and decided on 2850 relations. We conclude that the RT sub-classification was efficiently conducted and the method is both effective and practical.

1 INTRODUCTION

1.1 Japan Science and Technology Agency (JST) thesaurus

JST thesaurus being developed by JST is one of the largest scientific and technological thesauri. It contains 24.5 thousand concepts across a wide range of scientific and technological fields including the Life Sciences, Mechanics, Physics, Industrial Chemistry, Environmental Science, and Metallography. The JST thesaurus includes approximately 10,000 life-sciences concepts and its associated dictionary includes approximately 80,000 life-sciences concepts; some of which have links to MeSH (https://www.ncbi.nlm.nih.gov/mesh) terms. The concepts are structured using broader terms, narrower terms, and related terms (RT), and they are mainly used for the purpose of indexing scientific literature (http://jglobal.jst.go.jp/en/).

1.2 Background

To elucidate the mechanisms of biological phenomena, it is important to interpret what occurs at each level of molecules, cells, tissues, organs, and individuals and to define the relations among them. In this type of situation, specific biological ontologies, thesauri, and databases such as Gene Ontol-

ogy (http://www.geneontology.org/) are used for interpreting data from high-throughput experiments including Next-Generation Sequencing and microarray. These ontologies and databases have already proven to be essential knowledge bases for assisting in understanding mechanisms related to biological phenomena.

Although relations among such biological phenomena and gene products are being vigorously collected in Gene Ontology, there are almost no other ontologies, thesauri, or databases that have arranged relations among different categories and levels of biological phenomena such as the relationship between cellular and individual-level phenomena. One of the strong points of the JST thesaurus is that it widely collects information on the relations among biological concepts in different categories of the life-sciences field. For example, the JST thesaurus directly relates thromboembolism categorized into a disease to platelet aggregation categorized into a cellular phenomenon by using RT. In the thesaurus, information about concepts associated with RT is curated by experienced biological experts and can provide more reliable results than information based on co-occurrence among literature achieved by machine curation.

1.3 Refined JST thesaurus (ontology)

The original thesaurus of JST is mainly used for the purpose of indexing scientific literatures and extending retrieval terms, and it includes a wide range of terms, but the RTs among them are not rigorous. We aim to sub-classify the RTs and develop a new inter-linking ontology for biological concepts to solve the problems. It might become possible to describe more detailed and rigorous biological relations such as *gene products* that positively regulate *cellular physiological phenomena* to *disease*. We assign concepts of existing standard ontologies in the life-sciences field such as Semanticscience Integrated Ontology (http://sio.semanticscience.org) to the sub-classified relations to improve the versatility, reusability, and extensibility. Moreover, we plan to open the thesaurus to the public to perform an analysis of biological experimental data and
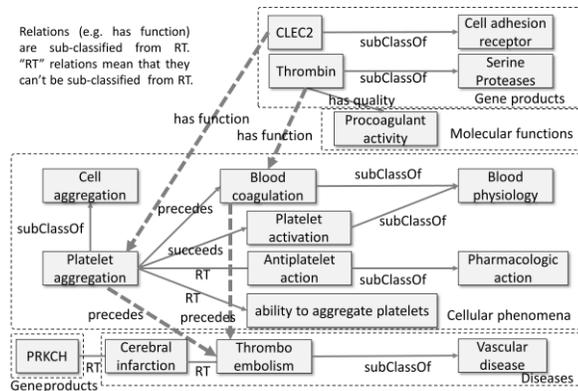

Fig. 1. Schema example of the developed ontology from JST thesaurus

assist in elucidating the mechanism of biological phenomena.

Furthermore, we will provide an information retrieval system in which the refined JST thesaurus (a new ontology) is implemented. This will not only allow researchers to investigate retrieval results from connections between concepts but also discover new knowledge according to inference or intelligent exploration. For example, by using the sub-classified relations such as the *has function* and *precedes* we can discover CLEC2 and thrombin as gene products strongly related to thromboembolism and exclude gene products distantly related to thromboembolism such as PRKCH that is connected to it through two RTs (Fig. 1 & Kushida *et al.*, 2016).

2   RELATED WORKS

Examples of the ontological development from thesauri and other language resources include YAGO. YAGO is constructed by unifying the categories and the infoboxes that are automatically extracted from Wikipedia with synsets of WordNet in a rule-based and heuristic method (Suchanek *et al.*, 2007). In Life Sciences, the examples include the conversion from thesaurus of agriculture and its related concepts (AGROVOC) into the ontology. In this project, the refining RT in more specific relation and the modeling using OWL are conducted (Soergel *et al.*, 2004).

Conversely, it is argued that merely specifying the relation of the thesaurus is insufficient for ontology construction (Kless *et al.*, 2016). Thus it is necessary to carefully design the structure of the relationship between concepts to convert the thesaurus into more a rigorous and solid ontology. In this study, by sub-classifying RT without defining the rigorous structure of concepts and the axiom, we aim to engineer a hybrid between thesaurus and ontology that has aspects of both forms. It is our future work to solve the differences between thesaurus and ontology and to explicitly define each of them, as pointed out by Kless *et al.* (2016).

The examples using the crowdsourcing include ontological alignment (Sarasua *et al.*, 2012) and the ontology's de-

velopment and maintenance (Mortensen et al., 2013). Mortensen *et al*. investigated crowdsourcing's performance for validating the relations among concepts in SNOMED CT (2015) and Gene Ontology (2016) and for validating the effects of the combination of crowdsourcing with medical experts' curation. LEGO (http://geneontology.org/page/connecting-annotations-lego-models) is an ongoing project where modeling semantic relations among biological processes, molecular functions, cellular components, and the related gene products is performed using expert crowdsourcing. The objective and approach are similar to that of the refining RT in JST thesaurus such as arranging biological relations by experts.

3   PAST APPROACH AND RESULTS

In this section, we summarize our past study (Kushida *et al*., 2016) and evaluate the validity of the method in detail.

3.1   Method of the sub-classification in 2016

We sub-classified 2065 RTs that made up approximately 42% (2065/4815) of all RTs in the life-sciences category in the JST thesaurus until March 2016. Four life-sciences experts including three curators were in charge of the practical implementation and sub-classification while one person (the manager) was in charge of the management and control of the sub-classification. The three curators had prior experience in indexing the JST thesaurus for scientific literature from 3 to >10 years although they were not experienced in handling ontologies. Conversely, the manager had experience in developing life-sciences ontologies.

The work was performed using the graphical ontology editor Hozo (http://www.hozo.jp/). The three curators sub-classified each RT to ten types of relations, namely, "subClassOf," "has part," "is part of," "has function," "is function of," "has quality," "is quality of," and "antonym" along with RT, following the guideline which had been created by the manager and had contained the definition of ten types of relations and the information of typical use examples.

When the three curators attempted to sub-classify an RT into a relation and it was agreed by three curators, we named it "the relation agreed by three curators (III-2016)." Likewise, when the sub-classified relation was agreed by two curators, we named it "the relation agreed by two curators (II-2016)."

Next, the manager confirmed whether each of the relations (III-2016 and II-2016) was correct or not. The following cases were used: case 1: when it was judged to be correct, the relation was determined as a result of the sub-classification; case 2: when it was judged to be incorrect, an appropriate relation was decided by the manager in consultation with the three curators; and case 3: when the relations which three curators had proposed were split (we named this situation as "Split-2016"), an appropriate relation was de-

cided by the manager in consultation with the three curators. We defined these relations decided by the process of the above three cases as "Correct relations." The number of relations of III-2016, II-2016, and Split-2016 was 1453 (70.4%), 580 (28.1%), and 32 (1.5%), respectively.

3.2   Method of evaluation

To quantitatively evaluate the validity of the RT sub-classifying method based on the majority decision, we calculated the precision and recall of III-2016 and II-2016.

Precision was calculated as the quotient of "the number of correct relations in the relations that were agreed by three or two curators (true positive)" divided by "the number of the relations that were agreed by three or two curators (true positive + false positive)."

Recall was calculated as the quotient of "the number of correct relations in the relations that were agreed by three or two curators (true positive)" divided by "the number of correct relations (true positive + false negative)." We only calculated recall for each relation but did not calculate recall for the sum of each relation. This is because in the case of the sum of relations, the denominator value of the calculation formula of recall will be equal to the denominator value of the precision; thus, the recall and precision values will be same. Therefore, we calculated the average of the recall for each relation instead of calculating the recall of the sum of each relation.

"Concentration rate" was defined to be an index of the degree of the answer tendency by curators. The concentration rate was calculated as the quotient of "the number of the relations that were agreed by three or two curators (true positive + false positive)" divided by "the number of correct relations (true positive + false negative)." These results were interpreted by an ontologist and life-science experts.

3.3   Error analysis of the III-2016 and II-2016

The precision of the sum of the relations in III-2016 (0.79) is higher than that in II-2016 (0.51) (Table 1). In III-2016 and II-2016, the precision of RT was somewhat low (0.78

Table 1. The performance of the RT sub-classification in III-2016, and II-2016

P, R, CR, and N stand for the precision, recall, concentration rate, and number of correct relations, respectively.

	III-2016				II-2016				All-2016		
Relation	P	R	CR	N	P	R	CR	N	P	R	CR
rdfs:subClassOf	0.88	0.25	0.28	60	0.83	0.46	0.56	123	0.84	0.39	0.47
skos:narrower (has narrower)	0.88	0.25	0.29	59	0.83	0.46	0.56	123	0.84	0.40	0.47
sio:SIO_000028 (has part)	1	0.04	0.04	53	0.83	0.11	0.14	44	0.88	0.07	0.08
sio:SIO_000068 (is part of)	1	0.04	0.04	54	0.83	0.11	0.14	44	0.88	0.07	0.08
sio:SIO_000218 (is quality of)	1	0.50	0.50	2	-	0	0	2	1	0.25	0.25
sio:SIO_000217 (has quality)	1	0.50	0.50	2	-	0	0	2	1	0.25	0.25
sio:SIO_000226(is function of)	1	0.23	0.23	73	0.95	0.44	0.47	45	0.97	0.31	0.32
sio:SIO_000225 (has function)	1	0.23	0.23	73	0.95	0.46	0.48	46	0.97	0.32	0.33
sio:SIO_000123 (antonym)	1	0.67	0.67	6	0.8	0.67	0.83	6	0.89	0.67	0.75
skos:related (RT)	0.78	1	1.28	1071	0.33	0.93	2.63	145	0.68	0.98	1.44
Sum (Mean) of relations	0.79 (0.37)		1.00	1453	0.51 (0.36)		1.00	580	0.71 (0.37)		1.00

rdfs: <http://www.w3.org/2000/01/rdf-schema#>, skos: <http://www.w3.org/2004/02/skos/core#>, sio: <http://semanticscience.org/resource/>

and 0.33) while the recall was high (1 and 0.93 respectively). Conversely, in III-2016 and II-2016, the precision of other relations except for RT such as "has part" (1 and 0.83) and "has function" (1 and 0.95) were high, and the recall of "has part" (0.04 and 0.11) and "has function" (0.23 and 0.46) were low. The concentration rate of RT in III-2016 and II-2016 were 1.28 and 2.63, respectively, and the values were more than that of other relations. This meant that the curator's answers seemed to be biased toward the RT.

Then, we examined the occurrence tendency of errors in III-2016 and II-2016 and observed that the total number of errors was 312 and 285, in which the number of errors relating to RT was 308 and 256, and the rate of errors relating to RT were 98.7% (308/312) and 89.8% (256/285), respectively. These results suggest that the three curators were unable to properly sub-classify RT into each relation such as "has part" and "has function." One reason for this might be that the curators did not fully understand the definitions and the usage of each relation. We considered that to solve this problem, it was necessary to revise the guideline for the sub-classification of RT to enhance curator training and to extend the graphical ontology editor Hozo as a curation tool.

3.4   Adding new relations as candidates for the RT sub-classifying

After finishing the sub-classification, in consultation with curators we decided to add 21 new relations as candidates, namely "synonym," "is connected to," "precedes," "succeeds," "has role," "is role of," "has phenotype," "is a Phenotype Of," "has output," "output of," "is similar to," "has creator," "is creator of," "has provider," "is provider of," "transforms into," "is transformed from," "is located in," "is location of," "regulate," and "is regulated by." We conducted the re-sub-classification using these 31 relations including the original ten relations (Kushida *et al.*, 2016).

4   IMPROVEMENT OF RT SUB-CLASSIFYING

Based on the results of the sub-classification conducted in 2016, we attempted to establish the process of efficiently developing an ontology from the JST thesaurus.

4.1   Revision of the guideline of RT sub-classifying and executing curators training

In addition to the definitions and usages of the 31 relations, we described the "domain" data that refers to the scope of a subject of each relation and the "range" that refers to the scope of an object of each relation in the guideline for the sub-classification of the RT (Table 2).

Furthermore, to fully understand each relation that is assigned in the sub-classification and to acquire basic knowledge about ontologies, the three curators participated in discussion about the creation of the guideline and have undertaken training in the sub-classification process using past data over a two-month period

4.2 Extending the ontology editor tool Hozo for RT sub-classifying

By accepting the proposals of curators, we improved the ontology editor Hozo to be able to input the first and second candidate relations. It was mandatory to input the first candidate relation and voluntary to input the second candidate relation.

Table 2. Examples of the domain and range of relations

Relation	Domain	Range
sio:SIO_000217 (has quality)	Continuant and Occurent	Quality
sio:SIO_000225 (has function)	Continuant (Material, Component, Localization, Organism)	Occurent (Event, Process, State)
sio:SIO_000228 (has role)	Continuant (Material)	Continuant (Material)
sio:SIO_001279 (has phenotype)	Occurent (Disease)	Occurent (Symptom)
sio:SIO_001154 (regulates)	Continuant (Cell, Material)	Occurent (Event, Process, State)
xkos:precedes	Occurent (Event, Process, State)	Occurent (Event, Process, State)

xkos: <http://rdf-vocabulary.ddialliance.org/xkos#>

5 METHOD OF SUB-CLASSIFICATION IN 2017

By following the new revised guidelines and using the extended Hozo tool, we sub-classified 2850 RT that was approximately 58% (2850/4815) of all of RT in the life-sciences category in JST thesaurus until March 2017. This was conducted by the same three trained curators and one manager as before.

When we sub-classified RT and in the first candidates a relation was agreed by three curators, we named the relation "1st-III." Likewise, we named the relation "1st-II:2nd-III," when in the first candidates, a relation was agreed by two curators, and in the first and second candidates, a relation was agreed by three curators. Moreover, when in the first candidates, a relation was agreed by two curators, and in the first and second candidates a relation was agreed by two curators, we named the relation "1st-II:2nd-II." In the case that the first candidate's relations proposed by the three curators were split and in the first and second candidates a relation was agreed by two curators, we named the relation "1st-Split:2nd-II." When in both of the first candidates and the second candidate's relations which the three curators proposed were split, we named the relation "1st-Split:2nd-Split" (Fig. 2).

Next, the manager confirmed whether each of the agreed relations (of 1st-III, 1st-II:2nd-III, 1st-II:2nd-II and 1st-Split:2nd-II) were correct or not, and case 1: when it was judged to be correct, the relation was determined as a result of the sub-classification, case 2: when it was judged to be incorrect, an appropriate relation was decided by the manager in consultation with three curators, and case 3: when the relations which three curators had proposed were split, namely 1st-Split:2nd-Split, an appropriate relation was decided by the manager in consultation with three curators likewise.

6 RESULTS OF SUB-CLASSIFICATION IN 2017

6.1 Result of 1st-III

Both the precision scores and the recall scores of each relation were high (see Table 3). In comparison with the results of III-2016 which corresponded to 1st-III, both of the precision and the recall of 1st-III was higher than those of III-2016, especially the recall of 1st-III was considerably higher than that of III-2016, e.g., "has function" and "has part" (Table 1 & 3).

Then, we investigated error occurrence. As a result, the total number of errors was 28 in which the number of errors relating to "subClassOf" was the most (18 errors, 64.3% (18/28)) of all. Examples included a relationship between "Carbon Cycle" and "biogeochemical cycle," and the correct relation was "is part of." Conversely, the number of errors relating to RT (6 errors, 21.4% (6/28)) that was the most in III-2016 (308 errors, 98.7% (308/312)) greatly decreased.

6.2 Result of 1st-II:2nd-III

The precision and the recall of each relation in 1st-II:2nd-III were as high as or slightly lower than that in 1st-III (Table 3). The total number of errors was ten, the number of errors relating to RT was the most (6 errors, 60.0% (6/10)) of all. Examples included a relationship between "Ascaride" and "parasite," and the correct relation was "has role."

6.3 Result of 1st-II:2nd-II

The precision of each relation was more than 0.83 except for "is similar to" (0.33 (2/6)), "has provider" (0 (0/1)), and "is provider of" (0 (0/1)) (Table 3). The recall of each relation was more than 0.89 except for "has creator" (0 (0/2)), "is creator of" (0 (0/2)), "has function" (0.63 (20/32)), "is function of" (0.64 (21/33)), "has role" (0.30 (10/57)), and "is role of" (0.30 (10/57)).

The total number of errors was 144 in which the number of errors relating to "RT" was the most (88 errors, 61.1% (88/144)) of all. Examples included the relationship between "Blattaria" and "insanitary insect" and the correct relation was "has role" and included the relationship between "body cavity camera" and "celioscopy" and the correct relation was "has function."

6.4 Results of 1st-Split:2nd-II

The precision of each relation was more than 0.80 except for "is location of" (0.33 (1/3)) and "is located in" (0.33 (1/3)) (Table 3). The recall of each relation was more than 0.75 except for "has creator" (0 (0/2)) and "is creator of" (0 (0/2)). The total number of errors was six in which the number of errors relating to "is location of" was the most (4 errors, 66.7% (4/6)) of all. Examples included a relationship

between "oil seed" and "Plant oils" and the correct relation was "is creator of."

6.5 Summary of the precision

We named the sum of 1st-III and 1st-II:2nd-III "III-2017," and we obtained the precision of III-2017 (0.97) and the average of recall (0.86) (Fig. 2). Moreover, we named the sum of 1st-II:2nd-II and 1st-Split:2nd-II "II-2017," and we obtained the precision score of II-2017 (0.87) and the average of recall (0.84). III-2017 meant the sum of the relations agreed in first and second candidates by three curators, and II-2017 meant the sum of relations agreed in first and second candidates by two curators. We summarize this as follows (Fig. 2 & Table 1),

- The precision of III-2017 (0.97) was higher than that of II-2017 (0.87).
- The precision of III-2016 (0.79) was higher than that of II-2016 (0.51) (Section 3.3 & Table 1).
- The precision of III-2017 (0.97) was higher than that of III-2016 (0.79).
- The precision of II-2017 (0.87) was higher than that of II-2016 (0.51).

As a result, we confirmed that the precision was improved by using the results of relations agreed by three curators such as III-2017 and III-2016, and the modified method in 2017 such as III-2017 and II-2017.

6.6 Summary of the recall

We compared the average recall of III-2017 and II-2017 with that of III-2016 and II-2016 and summarized the observations as follows (Fig. 2 & Table 1),
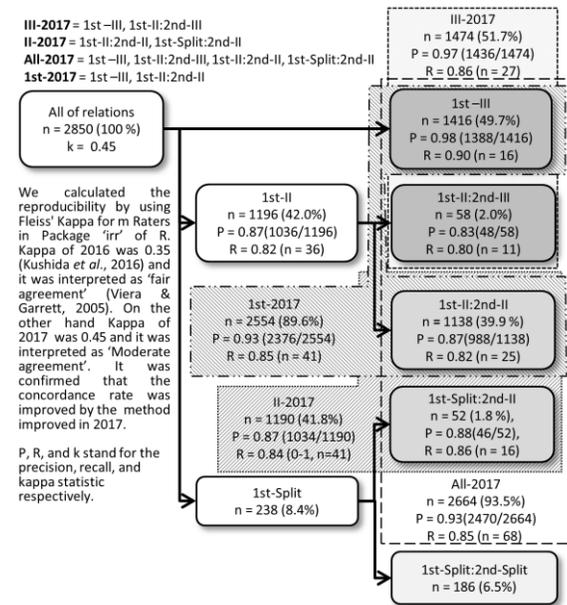
- The recall of III-2017 (0.86) was as much as that of II-2017 (0.84).
- The recall of III-2016 (0.37) was as much as that of II-2016 (0.36) (Table 1).
- The recall of III-2017 (0.86) was higher than that of III-2016 (0.37).
- The recall of II-2017 (0.84) was higher than that of II-2016 (0.36).

As a result, we confirmed that the recall was improved by using the modified method in 2017 such as III-2017 and II-2017. Conversely, we did not recognize that the recall would be considerably improved by using the information of the relations agreed by the three curators.

7 EVALUATION OF METHOD IN 2017

To evaluate effects of the sub-classification using the second candidate information, we compared the precision and recall of the sub-classification using both of the first and second relations information with that using the first candidate information only. We named relations agreed by two or three curators in both of the first and second candidates "All-2017." Namely, it included 1st-III, 1st-II:2nd-III, 1st-II:2nd-II, and 1st-Sprit:2nd-II (Fig. 2). We also named relations agreed by two or three curators in first candidate "1st-2017." Namely, it included 1st-III and 1st-II:2nd-II.

We calculated the precision and recall of All-2017 and 1st-2017. As a result, there was not much difference in the

III-2017 = 1st −III, 1st-II:2nd-III
II-2017 = 1st-II:2nd-II, 1st-Split:2nd-II
All-2017 = 1st −III, 1st-II:2nd-III, 1st-II:2nd-II, 1st-Split:2nd-II
1st-2017 = 1st −III, 1st-II:2nd-II

All of relations
n = 2850 (100 %)
k = 0.45

We calculated the reproducibility by using Fleiss' Kappa for m Raters in Package 'irr' of R. Kappa of 2016 was 0.35 (Kushida et al., 2016) and it was interpreted as 'fair agreement' (Viera & Garrett, 2005). On the other hand Kappa of 2017 was 0.45 and it was interpreted as 'Moderate agreement'. It was confirmed that the concordance rate was improved by the method improved in 2017.

P, R, and k stand for the precision, recall, and kappa statistic respectively.

1st-II
n = 1196 (42.0%)
P = 0.87(1036/1196)
R = 0.82 (n = 36)

1st-2017
n = 2554 (89.6%)
P = 0.93 (2376/2554)
R = 0.85 (n = 41)

II-2017
n = 1190 (41.8%)
P = 0.87 (1034/1190)
R = 0.84 (0-1, n=41)

1st-Split
n = 238 (8.4%)

III-2017
n = 1474 (51.7%)
P = 0.97 (1436/1474)
R = 0.86 (n = 27)

1st −III
n = 1416 (49.7%)
P = 0.98 (1388/1416)
R = 0.90 (n = 16)

1st-II:2nd-III
n = 58 (2.0%)
P = 0.83(48/58)
R = 0.80 (n = 11)

1st-II:2nd-II
n = 1138 (39.9 %)
P = 0.87(988/1138)
R = 0.82 (n = 25)

1st-Split:2nd-II
n = 52 (1.8 %),
P = 0.88(46/52),
R = 0.86 (n = 16)

All-2017
n = 2664 (93.5%)
P = 0.93(2470/2664)
R = 0.85 (n = 68)

1st-Split:2nd-Split
n = 186 (6.5%)

Fig. 2. The curators' agreement results for the first and second candidate relations in the RT sub-classification of 2017

Table 3. Performance of the RT sub-classification in 1st-III, 1st-II:2nd-III, 1st-II:2nd-II, and 1st-Split:2nd-II

Relation	1st-III				1st-II:2nd-III				1st-II:2nd-II				1st-Split:2nd-II			
	P	R	CR	N	P	R	CR	N	P	R	CR	N	P	R	CR	N
rdfs:subClassOf	0.95	1	1.06	161	0.93	1	1.08	13	0.83	0.99	1.19	104	0.80	1	1.25	4
skos:narrower (has narrower)	0.95	1	1.06	161	0.93	1	1.08	13	0.83	0.99	1.19	104	0.80	1	1.25	4
sio:SIO_000028 (has part)	1	0.74	0.74	23	-	-	-	0	1	0.89	0.89	18	1	1	1	7
sio:SIO_000068 (is part of)	1	0.74	0.74	23	-	-	-	0	1	0.89	0.89	18	1	1	1	7
sio:SIO_000218 (is quality of)	0.83	1	1.20	5	1	1	1	2	0.86	1	1.17	12	1	1	1	1
sio:SIO_000217 (has quality)	0.83	1	1.20	5	1	1	1	2	0.86	1	1.17	12	1	1	1	1
sio:SIO_000226 (is function of)	0.94	1	1.06	16	-	-	-	0	0.91	0.64	0.70	33	-	-	-	0
sio:SIO_000225 (has function)	0.94	1	1.06	16	-	-	-	0	0.91	0.63	0.69	32	-	-	-	0
sio:SIO_000123 (antonym)	1	1	1	14	-	-	-	0	1	1	1	4	-	-	-	0
sio:SIO_000228 (has role)	1	0.50	0.50	6	1	0.25	0.25	4	1	0.30	0.30	57	1	1	1	3
sio:SIO_000227 (is role of)	1	0.50	0.50	6	1	0.25	0.25	4	1	0.30	0.30	57	1	1	1	3
sio:SIO_001279 (has phenotype)	-	-	-	0	-	-	-	0	-	-	-	0	-	-	-	0
nbdc:isPhenotypeOf	-	-	-	0	-	-	-	0	-	-	-	0	-	-	-	0
sio:SIO_001154 (regulates)	-	-	-	0	-	-	-	0	1	1	1	2	-	-	-	0
sio:SIO_001155 (is regulated by)	-	-	-	0	-	-	-	0	1	1	1	2	-	-	-	0
xkos:succeeds	1	1	1	8	0.75	0.75	1	4	0.90	1	1.11	19	1	1	1	2
xkos:precedes	1	1	1	8	0.75	0.75	1	4	0.90	1	1.11	19	1	1	1	2
sio:SIO_000657 (is transformed from)	1	1	1	1	1	1	1	1	1	1	1	2	-	-	-	0
sio:SIO_000655 (transforms into)	1	1	1	1	1	1	1	1	1	1	1	2	-	-	-	0
sio:SIO_000203 (is connected to)	-	-	-	0	-	-	-	0	1	1	1	2	-	-	-	0
sio:SIO_000365 (is creator of)	-	-	-	0	-	-	-	0	-	0	0	2	-	0	0	2
sio:SIO_000364 (has creator)	-	-	-	0	-	-	-	0	-	0	0	2	-	0	0	2
sio:SIO_000145 (is location of)	-	-	-	0	-	-	-	0	1	1	1	2	0.33	1	3.00	1
sio:SIO_000061 (is located in)	-	-	-	0	-	-	-	0	1	1	1	2	0.33	1	3.00	1
obo:RO_0002234 (has output)	-	-	-	0	-	-	-	0	-	-	-	0	-	-	-	0
obo:RO_0002353 (output of)	-	-	-	0	-	-	-	0	-	-	-	0	-	-	-	0
sio:SIO_000064 (is provider of)	-	-	-	0	-	-	-	0	0	-	-	0	-	-	-	0
sio:SIO_000066 (has provider)	-	-	-	0	-	-	-	0	0	-	-	0	-	-	-	0
sio:SIO_000122 (synonym)	-	-	-	0	-	-	-	0	1	1	1	2	1	1	1	4
sio:SIO_000283 (is similar to)	-	-	-	0	-	-	-	0	0.33	1	3.00	2	-	-	-	0
skos:related (RT)	0.99	0.99	1	962	0.57	0.80	1.40	8	0.87	0.95	1.09	627	1	0.75	0.75	8
Sum (Mean) of relations	0.98 (0.90)	1.00		1416	0.83 (0.80)	1.00		58	0.87 (0.82)	1.00		1138	0.88 (0.86)	1.00		52

nbdc: <http://purl.jp/4/ontology/>, obo: <http://purl.obolibrary.org/obo/>

precision and the recall between All-2017 (P = 0.93, R = 0.85) and that of 1st-2017 (P = 0.93, R = 0.85). We did not recognize the advantage of using the second candidate information in either the precision or the recall.

Nevertheless, we confirmed that the number of relations which the three curators disagreed on in the first candidate (238 relations, we named it "1st-Split") was reduced to the number of 1st-Split:2nd-Split (186 relations) by the second candidate's information (Fig. 2). The process in which the manager suggested appropriate relations for those on which the three curators disagreed was laborious and time-consuming. Therefore, we considered that utilizing the second candidate's information to reduce the number of relations disagreed on by the three curators might be effective in terms of reducing this burden.

Furthermore, with the three curators, we examined whether inputting the second candidate relation in addition to the first candidate was difficult or not. We discovered that this was not much of a burden and allowing the curators to input the second candidate relation could contribute to reducing the time needed to narrow down the choice to just one candidate and to relieve work stress.

8   PUBLICATION

Currently, we are preparing to make a new ontology developed by using the improved method which is open to the public. Until the preparation is finished, the ontology and related results including the guidelines and the curation data are available for evaluation purposes only. If this is required, please contact the corresponding author. Furthermore, we are also currently examining the usage of one of the Creative Commons licenses for the publication; the public SPARQL endpoint is now being prepared and is planned for release. Moreover, we intend to submit the ontology to BioPortal (http://bioportal.bioontology.org/).

9   CONCLUSIONS

We described a method of efficiently constructing a new life-science ontology from an existing scientific and technological thesaurus by a small panel of experts, which comprised three curators and one manager. In the three cases within the RT sub-classification process, the manager confirmed whether each of the relations that had been agreed by the curators was correct or not. The following cases were used. In case 1, when it was judged to be correct, the relation was determined as a result of the sub-classification. In case 2, when it was judged to be incorrect, an appropriate relation was decided by the manager in consultation with three curators. Finally, in case 3, when the relations that the three curators had proposed were split, an appropriate relation was decided by the manager in consultation with three curators. However, we realize that cases 2 and 3 are laborious and time-consuming. We assume that it is important to reduce the number of cases 2 and 3 as much as possible to efficiently perform RT sub-classification. The rate of case 2 and 3 for 2016 and 2017 was 30.5% (629/2065) and 13.3% (380/2850), respectively. We confirmed that the proportion was reduced by less than half by performing the RT sub-classification following the revised guidelines in combination with the Hozo ontology editor operated by trained curators.

From our interviews and discussions with the curators, we realized that the domain and range information of each relation described in the guidelines were of practical use for appropriately selecting relations in the RT sub-classification. Specifically, this information could potentially be useful for curators lacking full experience in developing ontologies.

Although we consider that life-science experts need to sub-classify the relations among biological concepts, we will attempt to evaluate the validity of the crowdsourcing in the RT sub-classification process and the effect of cost reduction using crowdsourcing in our future research.

ACKNOWLEDGEMENTS

REFERENCES

Kless, D., Jansen, L., and Milton, S. (2016). A content-focused method for re-engineering thesauri into semantically adequate ontologies using OWL. Semantic Web, **7(5)**, 543-576.

Kushida, T., Masuda T., Tateisi Y., Watanabe K., Matsumura K., Kawamura T., Kozaki K., and Takagi T. (2016). Refining JST thesaurus and discussing the effectiveness in life science research. *Proc. of IESD2016*.

Mortensen, J.M., Minty, E.P., Januszyk, M., Sweeney, T.E., Rector, A.L., Noy, N.F., and Musen, M.A. (2015). Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. J *Am Med Inform Assoc.*, **22**, 640–8.

Mortensen, J.M., Musen, M.A., and Noy, N.F. (2013). Crowdsourcing the verification of relationships in biomedical ontologies. *AMIA Annu Symp Proc.* **16**, 1020–9.

Mortensen, J.M., Telis, N, Hughey, J.J, Fan-Minogue, H., Van Auken, K., Dumontier, M., and Musen, M.A. (2016). Is the crowd better as an assistant or a replacement in ontology engineering? An exploration through the lens of the Gene Ontology. *J Biomed Inform.* **60**, 199–209.

Sarasua, C., Simperl, E., and Noy, N.F. (2012). CrowdMAP: Crowdsourcing Ontology Alignment with Microtasks. 11th ISWC.

Soergel, D., Lauser B., Liang, A., Fisseha F., Keizer, J., and Katz, S. (2004). Reengineering Thesauri for New Applications: the AGROVOC Example. New Applications of KOS., **4(4)**, 1-26.

Suchanek F.M., Kasneci, G., and Weikum, G. (2007). YAGO—A Core of Semantic Knowledge Unifying WordNet and Wikipedia. *Proc 16th International Conference on WWW*. 697–706.

Viera, A.J., Garrett, J.M., (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5), 360–363.