**RDF入力手順マニュアル**

作成日：2021年12月吉日

作成者：浜口悠貴

目次

RDF直接入力手順・・・p. 3–12

用意したprefixについて・・・p.13–14

predicate表現探索について・・・p. 15–16

ICD10表記について・・・p. 17

CSV to RDF convert 手順・・・p. 18–26

今までのトラブル対応・・・p. 27–35

やり残したこと・・・p. 36–37

RDF直接入力手順

(0) 最初にprefixを用意する。

@prefix humandbs: <http://humandbs.biosciencedbc.jp/> .

@prefix dcterms: <http://purl.org/dc/terms/> .

@prefix dc: <http://purl.org/dc/elements/1.1/> .

@prefix foaf: <http://xmlns.com/foaf/0.1/> .

@prefix owl: <http://www.w3.org/2002/07/owl#> .

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

@prefix terms: <http://med2rdf.org/gwascatalog/terms/> .

@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

@prefix pav: <http://purl.org/pav/>.

@prefix obo: <http://purl.obolibrary.org/obo/> .

@prefix OBI: <http://purl.obolibrary.org/obo/OBI\_> .

@prefix bao: <http://www.bioassayontology.org/bao#> .

@prefix efo: <http://www.ebi.ac.uk/efo/> .

@prefix edam: <http://edamontology.org/> .

@prefix mesh: <http://purl.bioontology.org/ontology/MESH/> .

@prefix Thesaurus: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#> .

@prefix icd\_10: <http://purl.bioontology.org/ontology/ICD10/> .

@prefix iobc: <http://purl.jp/bio/4/id/> .

@prefix snmi: <http://purl.bioontology.org/ontology/SNMI/> .

@prefix prism: <http://prismstandard.org/namespaces/basic/3.0/> .

@prefix bibo: <http://purl.org/ontology/bibo/> .

@prefix JGA: <http://ddbj.nig.ac.jp/resource/jga-dataset/> .

@prefix DRA: <http://trace.ddbj.nig.ac.jp/DRASearch/submission?acc=>.

@prefix GEA: <http://ddbj.nig.ac.jp/public/ddbj\_database/gea/experiment/E-GEAD-000/> .

@prefix JGA\_STUDY: <http://ddbj.nig.ac.jp/resource/jga-study/> .

@prefix EGA: <http://ega-archive.org/datasets/>.

@prefix EGA\_STUDY: <http://ega-archive.org/studies/>.

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

@prefix GBOL: <http://gbol.life/0.1/> .

@prefix org: <http://www.w3.org/ns/org#> .

@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

@prefix sb: <http://sparqlbook.jp/book/> .

@prefix dbo: <http://dbpedia.org/ontology/> .

@prefix MedGen\_UID: <http://www.ncbi.nlm.nih.gov/medgen/?term=> .

@prefix opmi: <http://purl.obolibrary.org/obo/opmi.owl> .

@prefix GAZ: <http://purl.obolibrary.org/obo/GAZ\_> .

(1) humandbsの後にNBDC Research ID(hum No.)を追加する。

humandbs:hum...

(2) dcterms:titleの後に研究題目を追加する。

dcterms:title "";

(3) 提供者の名前を追加する。

predicateは「dcterms:contributor」, 「a foaf」,「foaf:firstName」, 「foaf:lastName」を使う。

a foafの後にはPersonを、foaf:nameの後には提供者の名前を、foaf:nameの後には提供者の名、 foaf:lastNameの後には提供者の氏を追加する。

dcterms:contributor [

a foaf:Person;

foaf:name "Taro Yamada";

foaf:firstName "Taro";

foaf:lastName "Yamada"

];

(4) 参考文献とデータIDとdoiを追加する。

predicateは「dcterms:references」「dc:title」「rdfs:resource」「 terms:study\_id」「 dcterms:identifier」「prism:doi」「prism:url」を使う。

dc:titleには論文のタイトルを、rdfs:resourceには論文に対応するデータ情報\*を、

dcterms:identifierにはdoi:doi番号をprism:doiにはdoi番号をprism:urlにはdoiのurlを入力する

(\*注意点：データ IDは外部のデータID(JGAD...とDRA...)と内部のデータID(hum...など)に分かれており、それぞれ記載方法が違う。rdfs:resource JGAの場合はJGAD番号を rdfs:resource DRAの場合にはDRA番号を、rdfs:resource以下のterms:study\_idにhum番号を入力すること）

dcterms:references [

dc:title "XXX.";

rdfs:resource [

terms:study\_id "hum...."

];

rdfs:resource JGA:JGAD....;

rdfs:resource DRA:DRA...;

dcterms:identifier "doi:ZZZ";

prism:doi "ZZZ" ;

prism:url <http://dx.doi.org/ZZZ>

];

(5) 日付とversionを追加する。

predicateは「pav:version」「owl:versionInfo」「 dcterms:modified」を使う。

owl:versionInfoの後にversion No. \*、dcterms:modifiedの後に""^^xsd:dateつきで公開日を入力する。

(\*注意点：version No.は数字のみ入力。""は不要)

pav:version [

owl:versionInfo X;

dcterms:modified "XXXX/XX/XX"^^xsd:date;

(6) dataset情報を入力する。

predicateは「terms:dataset」「 terms:dataset\_member」「 terms:data\_type」を使う。

datasetごとに詳細情報を記載していく\*。terms:dataset\_member以下の

terms:study\_idの後にデータID\*\*、terms:data\_typeの後にアクセス制限情報\*\*\*を入力する。

(\*注意点1：続く(7)~(9)の情報は 各terms:datasetごとに続けて入力していくこと)

(\*\*\*注意点2：terms:study\_idに入るデータがJGAのstudy\_idの場合はJGA\_STUDY:JGAS....と記載している。(4)のrdfs:resource JGA:JGAD....と混同しないように。)

(\*\*\*注意点3：アクセス制限情報の表記ゆれを防ぐために"Unrestricted-Access"と"Controlled-Access (Type I)"と"Controlled-Access (Type II)"の書き方のみを推奨。)

terms:dataset [

terms:dataset\_member [

terms:study\_id "XXXX";

terms:data\_type "Unrestricted-Access";

];

terms:dataset\_member [

terms:study\_id JGA\_STUDY:JGAS....;

terms:data\_type "Controlled-Access (Type II)";

];

terms:dataset\_member [

terms:study\_id DRA:JGA....;

terms:data\_type "Unrestricted-Access";

];

(7) データの種類(Type of study)を入力する。

predicateは「terms:technology」「 rdfs:label」「 terms:technology\_ontology」「terms:JGA\_ID」「terms:DRA\_ID」「terms:DATA\_DIRECTORY 」「terms:NBDC\_DATA」「efo:EFO\_0000651」「terms:phenotype\_file」「terms:result\_file」を使う。

rdfs:label の後にはデータの種類を、terms:technology\_ontologyにはデータの種類のオントロジーを、「terms:JGA\_ID」「terms:DRA\_ID」「terms:NBDC\_DATA」以下に対応するデータ番号を入力する。ただし、phenotype情報と解析・統計情報の2情報を記載する必要がある場合は「terms:DATA\_DIRECTORY 」以下に入力する。「efo:EFO\_0000651」の後にはphenotype情報を、 terms:phenotype\_fileにはphenotype情報に対応するデータ

を、terms:result\_fileには解析・統計情報を入力する。

terms:technology [

rdfs:label "XXX";

terms:technology\_ontology YYY;

terms:JGA\_ID JGA:JGA....;

terms:NBDC\_DATA <ZZZ>;

terms:DATA\_DIRECTORY [

efo:EFO\_0000651 "...";

terms:phenotype\_file [

terms:JGA\_ID JGA:JGAD000101;

];

terms:result\_file [

terms:NBDC\_DATA <...>;

];

];

(8) 参加者(対象集団)を入力する。

predicateは「terms:participant」「rdfs:label」「terms:participant\_member」「terms:status 」「terms:individuals」「terms:sample\_number」「 terms:disease\_id」「terms:NBDC\_DATA」「terms:JGA\_ID」を使う。

rdfs:label の後には全患者情報を、terms:statusの後には各患者名を、「opmi:owl

」の後には出身(国籍または地域)や性別の情報、「terms:individuals」の後には患者数(組数)\*と「terms:sample\_number」の後には患者数(合計)\*を、 terms:disease\_idにはICD10情報またはMedGen\_UID情報を、「terms:JGA\_ID」「terms:DRA\_ID」「terms:DATA\_DIRECTORY 」「terms:NBDC\_DATA」以下に各患者情報に対応するデータ番号を入力する。

(\*注意点1：患者数(組数)、患者数(合計)ともに数字のみを入力すること)

terms:participant [

rdfs:label "XXX";

opmi:owl "...";

terms:participant\_member [

terms:status "YYY";

terms:individuals "";

terms:sample\_number 9;

terms:disease\_id icd10:....;

terms:NBDC\_DATA <...>;

];

terms:participant\_member [

terms:status "ZZZ";

terms:individuals 6;

terms:sample\_number 12;

terms:disease\_id "";

terms:JGA\_ID JGA:...;

];

];

(9) 手法情報を入力する

predicateは「terms:machine」「a」「terms:technology\_ontology」「 rdfs:label」「terms:machine\_name\_ontology」「OBI:0000304」「terms:company\_ontology」を使う。

aの後には手法情報を、terms:technology\_ontologyの後には手法情報のオントロジーを、rdfs:labelの後には機器名を、terms:machine\_name\_ontologyの後には機器情報のオントロジーを、OBI:0000304 以下のrdfs:labelの後には会社名、 OBI:0000304 以下のterms:machine\_name\_ontologyには会社のオントロジー情報を、terms:device以下のrdfs:labelの後にはkitまたはchip名を、terms:device以下の terms:device\_ontologyには

試薬・チップのオントロジーを追加する。

terms:machine [

a "XXX";

terms:technology\_ontology XXX;

rdfs:label "XXX";

terms:machine\_name\_ontology "XXX";

OBI:0000304 [

rdfs:label "XXX";

terms:company\_ontology XXX;

];

terms:device [

rdfs:label "XXX";

terms:device\_ontology XXX;

];

(10) ttl文を終結する。以下の3行を入力する

];

];

].

(11)ttl文が間違っていないか確認する。Turtle Validator (http://ttl.summerofcode.be)を使って確認。

# RDF入力例

## ①humNo.232

humandbs:hum0232

dcterms:title "Evaluation of metastasis and recurrence by monitoring circulating tumor DNA of gastric cancer";

dcterms:contributor [

a foaf:Person;

foaf:name "Satsoshi Nishizuka";

foaf:firstName "Satsoshi";

foaf:lastName "Nishizuka"

];

dcterms:references [

dc:title "Analysis of mutational and proteomic heterogeneity of gastric cancer suggests an effective pipeline to monitor post-treatment tumor burden using circulating tumor DNA";

rdfs:resource JGA:JGAD000325 ;

dcterms:identifier "doi:10.1371/journal.pone.0239966";

prism:doi "10.1371/journal.pone.0239966" ;

prism:url "http://dx.doi.org/10.1371/journal.pone.0239966"

];

pav:version [

owl:versionInfo 1;

dcterms:modified "2020/6/1"^^xsd:date;

terms:dataset [

terms:dataset\_member [

terms:study\_id JGA\_STUDY:JGAS000231;

terms:data\_type "Controlled-Access (Type I)";

];

terms:technology [

rdfs:label "Target Capture";

terms:technology\_ontology OBI:0001899;

terms:JGA\_ID JGA:JGAD000325 ;

terms:participant [

rdfs:label "Gastric cancer (ICD10: C169): 10 cases, primary tumor tissue: 30 samples, PBMC: 10 samples";

terms:participant\_member [

terms:status "Gastric cancer (primary tumor tissue)";

terms:individuals 10;

terms:sample\_number 30;

terms:disease\_id icd\_10:C16.9;

];

terms:participant\_member [

terms:status "Gastric cancer (PBMC)";

terms:individuals 10;

terms:sample\_number 10;

terms:disease\_id icd\_10:C16.9;

];

];

terms:machine [

a "high throughput sequencer";

terms:machine\_ontology efo:EFO\_0002702;

rdfs:label "HiSeq 2000";

terms:machine\_name\_ontology OBI:0002001 ;

OBI:0000304 [

rdfs:label "illumina";

terms:company\_ontology OBI:0000759

];

terms:device [

rdfs:label "ClearSeq SS Comprehensive Cancer kits";

terms:device\_ontology ""

]

]

]

]

].

## ② humNo.194

humandbs:hum0194

dcterms:title "Molecular mechanisms of drug resistance in cancer";

dcterms:contributor [

a foaf:Person;

foaf:name "Ryohei Katayama";

foaf:firstName "Ryohei";

foaf:lastName "Katayama"

];

dcterms:references [

dc:title "The new-generation selective ROS1/NTRK inhibitor DS-6051b overcomes crizotinib resistant ROS1-G2032R mutation in preclinical models";

rdfs:resource JGA:JGAD000272 ;

dcterms:identifier "doi:10.1038/s41467-019-11496-z";

prism:doi "10.1038/s41467-019-11496-z" ;

prism:url <http://dx.doi.org/10.1038/s41467-019-11496-z>

];

pav:version [

owl:versionInfo 1;

dcterms:modified "2020/5/11"^^xsd:date;

terms:dataset [

terms:dataset\_member [

terms:study\_id JGA\_STUDY:JGAS000189;

terms:data\_type "Controlled-Access (Type I)";

];

terms:technology [

rdfs:label "Target Capture";

terms:technology\_ontology OBI:0001899;

terms:JGA\_ID JGA:JGAD000272;

terms:participant [

rdfs:label "Lung adenocarcinoma (ICD10: C34): 2 cases";

terms:participant\_member [

terms:status "Lung adenocarcinoma";

terms:individuals "";

terms:sample\_number 2;

terms:disease\_id icd\_10:C34;

]

];

terms:machine [

a "high throughput sequencer";

terms:machine\_ontology efo:EFO\_0002699;

rdfs:label "MiSeq";

terms:machine\_name\_ontology OBI:0002003 ;

OBI:0000304 [

rdfs:label "illumina";

terms:company\_ontology OBI:0000759

];

terms:device [

rdfs:label "Agilent HaloPlex HS Custom Kits";

terms:device\_ontology ""

]

]

]

]

].

## ③humNo.0166

humandbs:hum0166

dcterms:title "Multicenter Evaluation of NUDT15-R139C genotyping efficiency to DEtect thiopurine-induced hair Loss and leukopenia (MENDEL study)";

dcterms:contributor [

a foaf:Person;

foaf:name "Yoichi Kakuta";

foaf:firstName "Yoichi";

foaf:lastName "Kakuta"

];

dcterms:references [

dc:title "NUDT15 codon 139 is the best pharmacogenetic marker for predicting thiopurine-induced severe adverse events in Japanese patients with inflammatory bowel disease: a multicenter study";

rdfs:resource [

terms:study\_id "hum0166.v1.gwas.v1"

];

dcterms:identifier "doi:10.1007/s00535-018-1486-7";

prism:doi "10.1007/s00535-018-1486-7" ;

prism:url <http://dx.doi.org/10.1007/s00535-018-1486-7>

];

pav:version [

owl:versionInfo 1;

dcterms:modified "2019/7/30"^^xsd:date;

terms:dataset [

terms:dataset\_member [

terms:study\_id "hum0166.v1.gwas.v1";

terms:data\_type "Unrestricted-Access";

];

terms:technology [

rdfs:label "genome wide SNVs";

terms:technology\_ontology efo:EFO\_0002764;

terms:NBDC\_DATA <http://humandbs.biosciencedbc.jp/files/hum0166/hum0166.v1.gwas.v1.zip>;

terms:NBDC\_DATA <http://humandbs.biosciencedbc.jp/files/hum0166/hum0166.v1.gwas.v1\_header.xlsx>;

terms:participant [

rdfs:label "inflammatory bowel diseases: 1221 cases";

terms:participant\_member [

terms:status "Crohn's disease";

terms:individuals "";

terms:sample\_number 516;

terms:disease\_id icd\_10:K50.9

];

terms:participant\_member [

terms:status "Ulcerative Colitis";

terms:individuals "";

terms:sample\_number 674;

terms:disease\_id icd\_10:K51.9

];

terms:participant\_member [

terms:status "Gastrointestinal Behçet's disease";

terms:individuals "";

terms:sample\_number 16;

terms:disease\_id icd\_10:M35.2

];

terms:participant\_member [

terms:status " other inflammatory bowel diseases";

terms:individuals "";

terms:sample\_number 15;

terms:disease\_id icd\_10:K52.3

]

];

terms:machine [

a "microarray";

terms:machine\_ontology OBI:0400147;

rdfs:label "";

terms:machine\_name\_ontology "";

OBI:0000304 [

rdfs:label "Affymetrix";

terms:company\_ontology Thesaurus:C123894

];

terms:device [

rdfs:label "Axiom 2.0 Reagent Kit";

terms:device\_ontology ""

]

]

]

]

].

用意したprefixについて

(1)RDFで書くに当たって必須だったもの

@prefix humandbs: <http://humandbs.biosciencedbc.jp/> .

@prefix dcterms: <http://purl.org/dc/terms/> .

@prefix dc: <http://purl.org/dc/elements/1.1/> .

@prefix foaf: <http://xmlns.com/foaf/0.1/> .

@prefix owl: <http://www.w3.org/2002/07/owl#> .

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

@prefix org: <http://www.w3.org/ns/org#> .

@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

@prefix sb: <http://sparqlbook.jp/book/> .

@prefix dbo: <http://dbpedia.org/ontology/> .

(2)version情報

@prefix pav: <<http://purl.org/pav/>>.

(3)文献情報

@prefix prism: <http://prismstandard.org/namespaces/basic/3.0/> .

(4)データセット情報

@prefix JGA: <http://ddbj.nig.ac.jp/resource/jga-dataset/> .

@prefix DRA: <http://trace.ddbj.nig.ac.jp/DRASearch/submission?acc=>.

@prefix GEA: <http://ddbj.nig.ac.jp/public/ddbj\_database/gea/experiment/E-GEAD-000/> .

@prefix JGA\_STUDY: <http://ddbj.nig.ac.jp/resource/jga-study/> .

@prefix EGA: <http://ega-archive.org/datasets/>.

@prefix EGA\_STUDY: <<http://ega-archive.org/studies/>>.

(5)病名情報

@prefix icd\_10: <http://purl.bioontology.org/ontology/ICD10/> .

@prefix MedGen\_UID: <http://www.ncbi.nlm.nih.gov/medgen/?term=> .

(6)出身(国籍または地域)や性別の情報

@prefix Thesaurus: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#> .

@prefix opmi: <http://purl.obolibrary.org/obo/opmi.owl> .

@prefix GAZ: <http://purl.obolibrary.org/obo/GAZ\_> .

(7)手法や機器・試薬などの情報

@prefix Thesaurus: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#> .

@prefix obo: <http://purl.obolibrary.org/obo/> .

@prefix OBI: <http://purl.obolibrary.org/obo/OBI\_> .

@prefix bao: <http://www.bioassayontology.org/bao#> .

@prefix efo: <http://www.ebi.ac.uk/efo/> .

@prefix edam: <http://edamontology.org/> .

@prefix mesh: <http://purl.bioontology.org/ontology/MESH/> .

@prefix iobc: <http://purl.jp/bio/4/id/> .

@prefix snmi: <http://purl.bioontology.org/ontology/SNMI/> .

@prefix bibo: <http://purl.org/ontology/bibo/> .

@prefix GBOL: <http://gbol.life/0.1/> .

(8)自分で新しくpredicateを作るとき

@prefix terms: <http://med2rdf.org/gwascatalog/terms/> .

predicate表現探索について

# (1)オントロジー検索に有用なサイト

用意したprefixについて(7)に書いてあるprefixでオントロジーを探すのが主だと思われるが、オントロジー検索にとても有用なサイトは以下である。

①<http://www.ontobee.org>

大規模なオントロジー検索ならばこれ

②<https://bioportal.bioontology.org>

生物系の単語・オントロジー検索ならばこれ (例:illumina の機器や手法)

③<https://www.ebi.ac.uk/efo/>

実験手法や器具はなるべくEFO(The Experimental Factor Ontology)で統一したいとき向け

# (2)RDF表現を書くに当たって参考にしたサイト

最初はSQLに触ったぐらいの初心者である筆者がRDFをかくに当たって参考にしたサイトは以下である。

①TogoWiki

Guideline (<https://wiki.lifesciencedb.jp/mw/RDFizingDatabaseGuideline>)

ICD10 (<https://wiki.lifesciencedb.jp/mw/DiseaseID>)

②TogoTV

RDF作成のデザインパターン(<https://togotv.dbcls.jp/20191210.html>)

virtuosoインストール(<https://togotv.dbcls.jp/20191212.html>)

③endpoint-browser

(<https://sparql-support.dbcls.jp/endpoint-browser.html>)

④NBDC RDF portal

(<http://integbio.jp/rdf/datasets>)

# (3)自分は使わなかったが今後有用になるかもしれないサイト

①統計LOD (<https://data.e-stat.go.jp/lodw/>)

患者情報(出身など)の入力に役に立つかもしれない

②​​<https://rml.io/specs/rml/>

RDF Mapping Languageの説明サイト

ICD10表記について

[標準病名マスター病名検索](http://www.byomei.org/Scripts/search/index_search.asp) にて病名情報を探すのが主流だが、

RDFに追加する際に注意点が二つある。

(1)数字表記

例えば黒色腫と検索すると以下のように表示される。



アルファベットの後に数字が2桁または3桁続いているのが見える。3桁の場合に注意が必要である。RDFで書くとき3桁は「2桁」+「.」+「1桁」で表記する必要がある\*。

(\*RDFにはICD10情報として「http://purl.bioontology.org/ontology/ICD10/」のprefixを利用しているのだが、例えばC300は「http://purl.bioontology.org/ontology/ICD10/C300」

と検索しても出てこない。「http://purl.bioontology.org/ontology/ICD10/C30.0」が正解である。)

(2)総合表記

色々な病気をひっくるめて書くとき、?XX-?XXと表記するケースが出てくるが、その際も(1)同様に小数点に気を付けなければいけない。

誤：C00-D48　　　→正：C00-D48.9

誤：I20-I25　　　 →正：I20-I25.9

誤：C81-C96 →正：C81-C96.9

CSV to RDF convert 手順

hum0249を例に作成手順をここにしるす。[RDF\_format\_example.xlsx](https://docs.google.com/spreadsheets/d/1oIRMNgcuuEY1bMl2tbN0y7mZfjhudFJb/edit?usp=sharing&ouid=109913611248719322065&rtpof=true&sd=true)を参考にすること

＜hum番号+提供者情報＞

hum番号+提供者情報の欄をttlファイル上に貼り付ける。



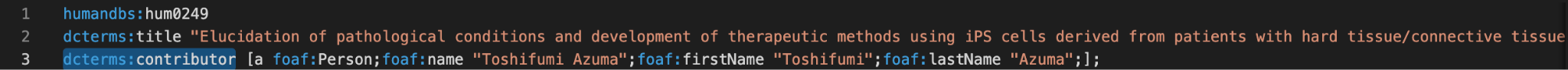
\tを空白に変換する



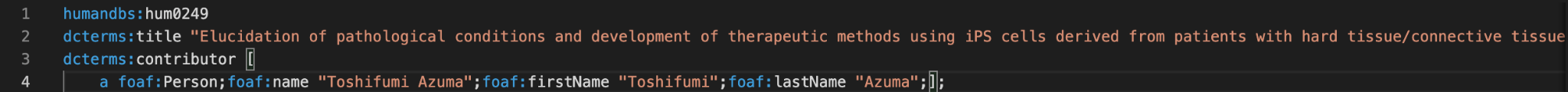
dcterms:titleの前(hum番号の後)に改行を入れる



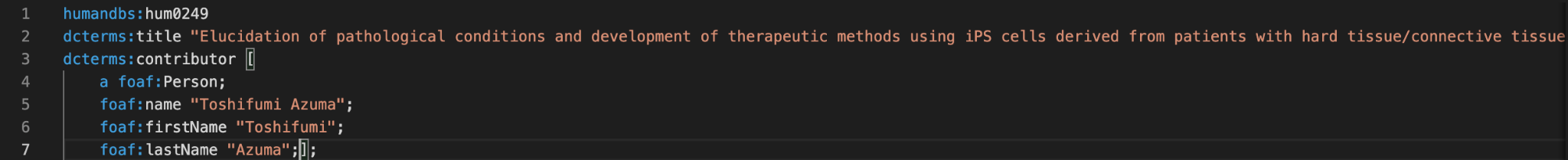
dcterms:contributorの前に改行を入れる



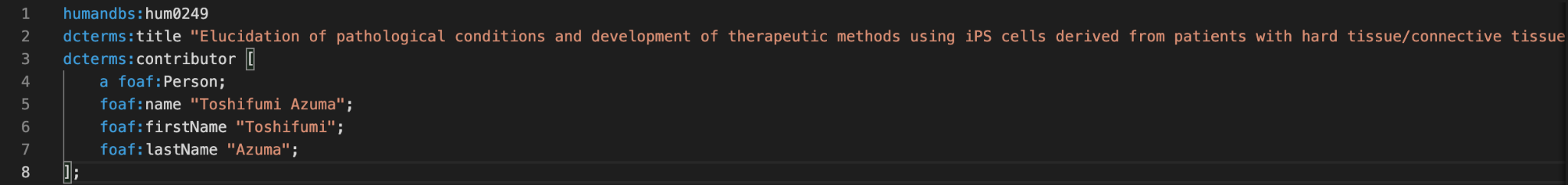
a foaf:Personの前に改行を入れる



foaf:nameとfoaf:firstNameとfoaf:lastNameの前に改行を入れる



];の位置を調整



# 

# ＜参考文献＞

参考文献の欄をttlファイル上に貼り付ける。



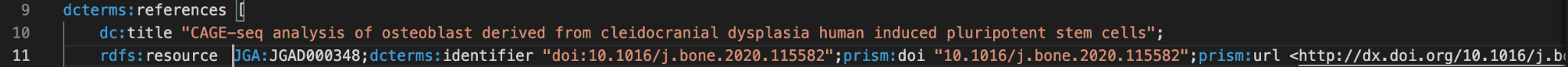
\tを空白に変換する



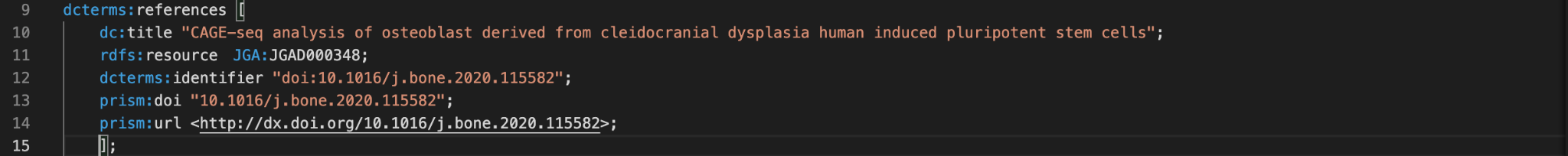
dcterms:references [の後に改行を追加する



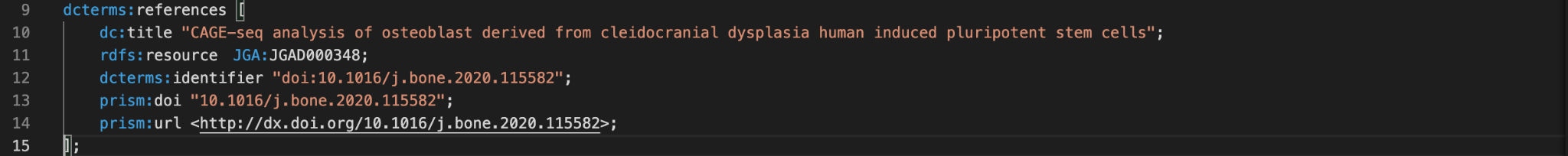
rdfs:resourceの前に改行を追加し、後に空文字を入れる



];の前の;後に改行を追加していく



微調整



# ＜version番号+dataset名＞

version番号+dataset名の欄をttlファイル上に貼り付ける。



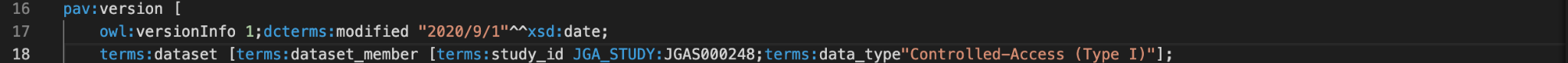
\tを空白に変換する



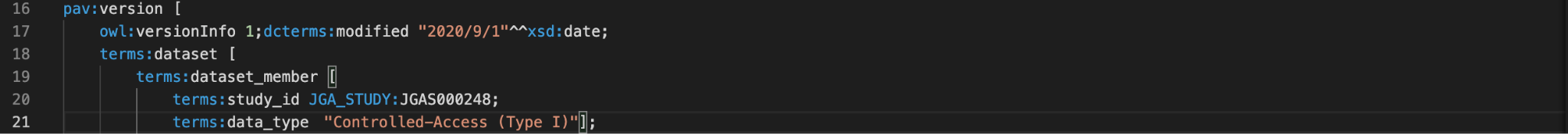
pav:version [の後に改行を追加する



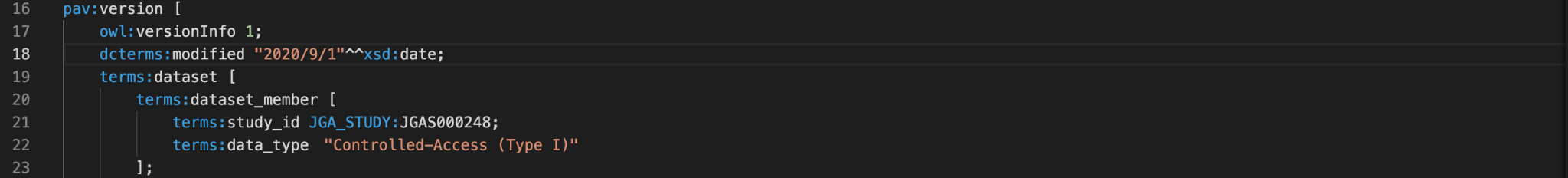
owl:versionInfoの後に空文字を入れ、terms:dataset　[の前に改行を追加する



terms:dataset [　の後、terms:dataset\_member [　の後、terms:data\_typeの前に改行を追加し、terms:data\_typeの後に空文字を入れる



微調整



# ＜Technology＞

technologyの欄をttlファイル上に貼り付ける。



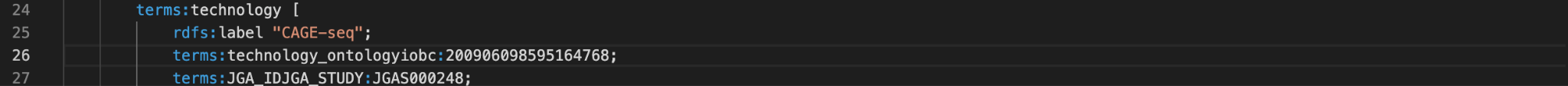
\tを空白に変換する

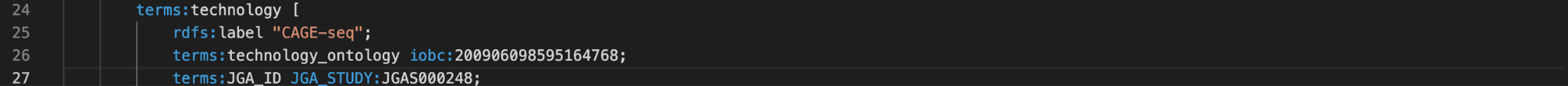


terms:technology [の前に\t\tを挿入



terms:technology [の;後に改行を追加

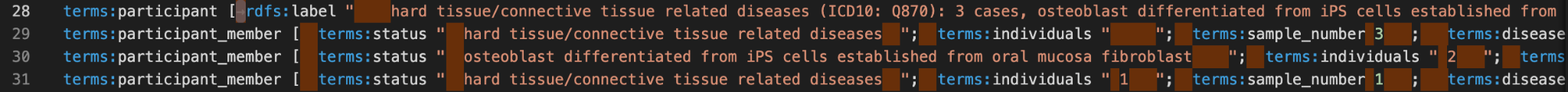


terms:technology\_ontologyとterms:JGA\_IDの後に空文字を追加する

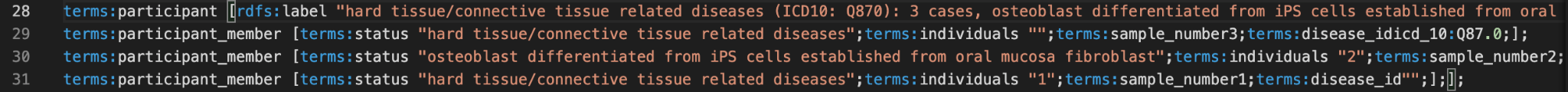
# 

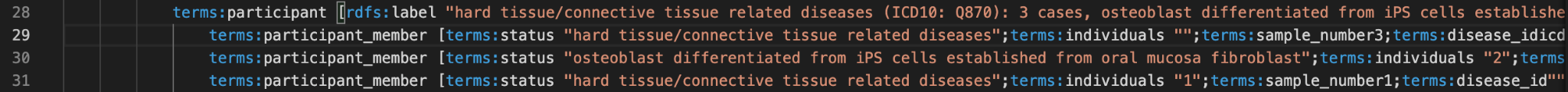
# ＜患者情報＞

患者情報の欄をttlファイル上に貼り付ける。

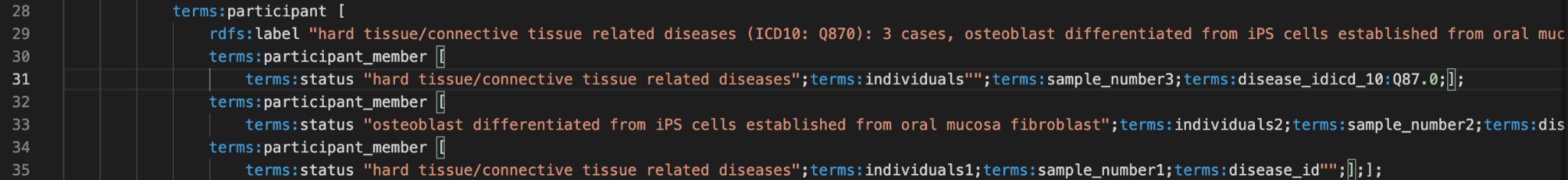


\tを空白に変換する

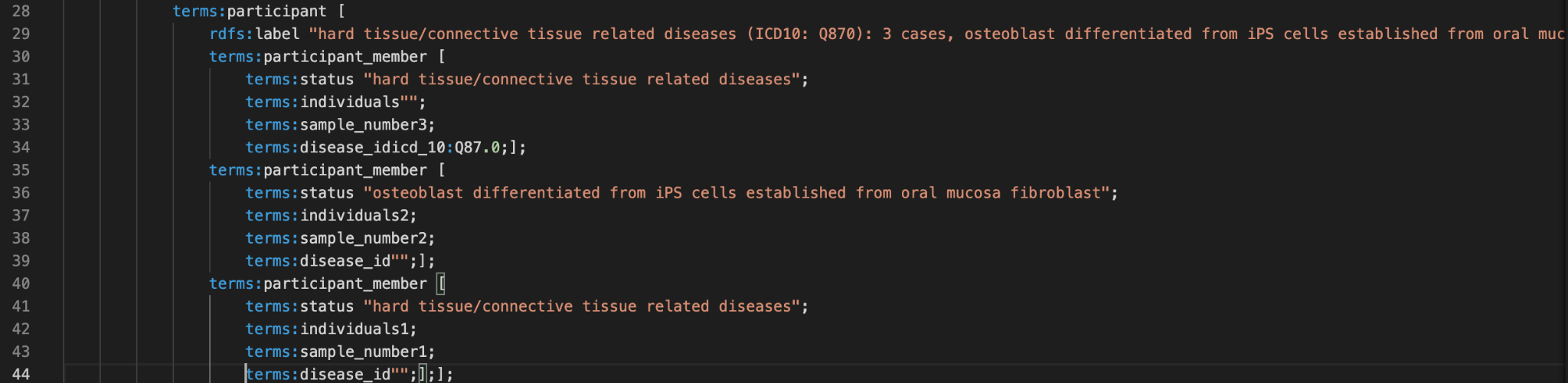
terms:participant [ の前に\tを3つ、terms:participant\_memberの前に\tを4つ入れる



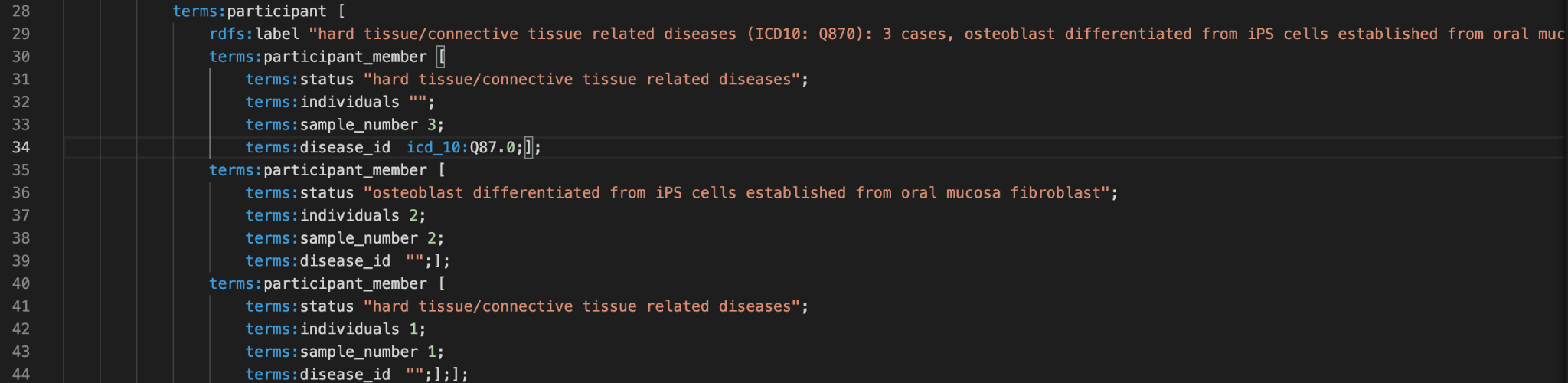
terms:participant [ とterms:participant\_memberの後に改行を入れる



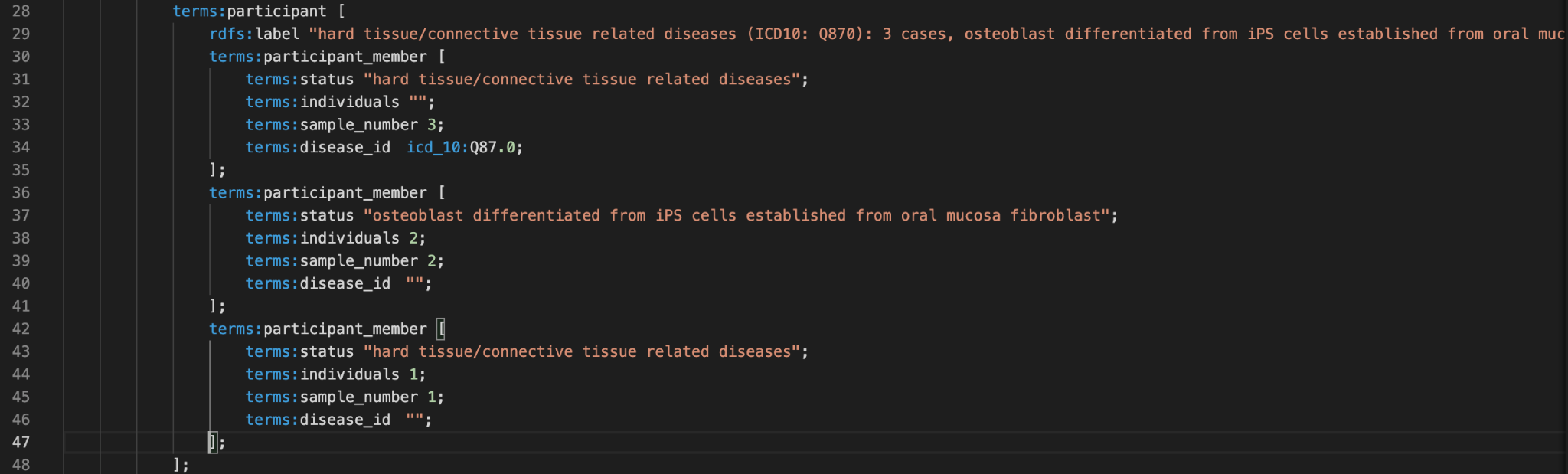
terms:individuals、terms:sample\_number、terms:disease\_idの前に改行を入れる



terms:individuals、terms:sample\_number、terms:disease\_idの後に空文字を追加する



微調整



# ＜Machine情報＞

機械情報の欄をttlファイル上に貼り付ける。



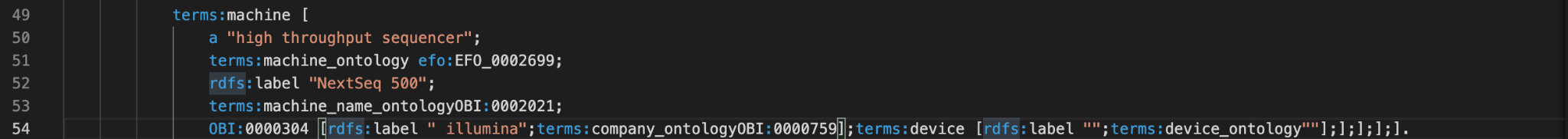
\tを空白に変換する



terms:machine [の前に\tを3つ挿入する



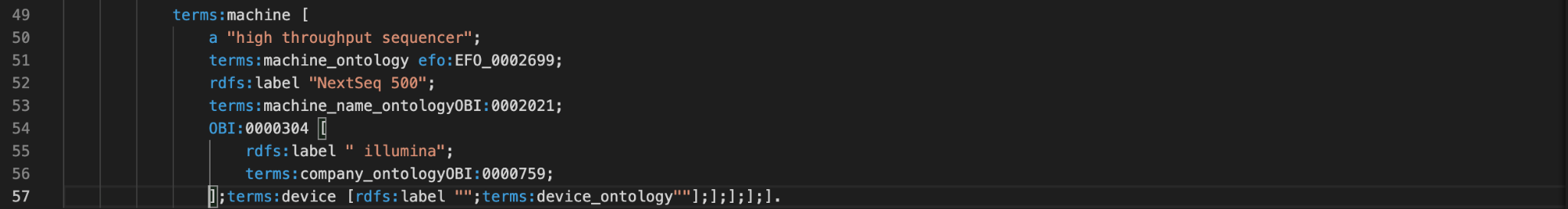
terms:machineの後とterms:machine\_ontology, rdfs:label, terms:machine\_name\_ontology, OBI:0000304の前に改行を追加



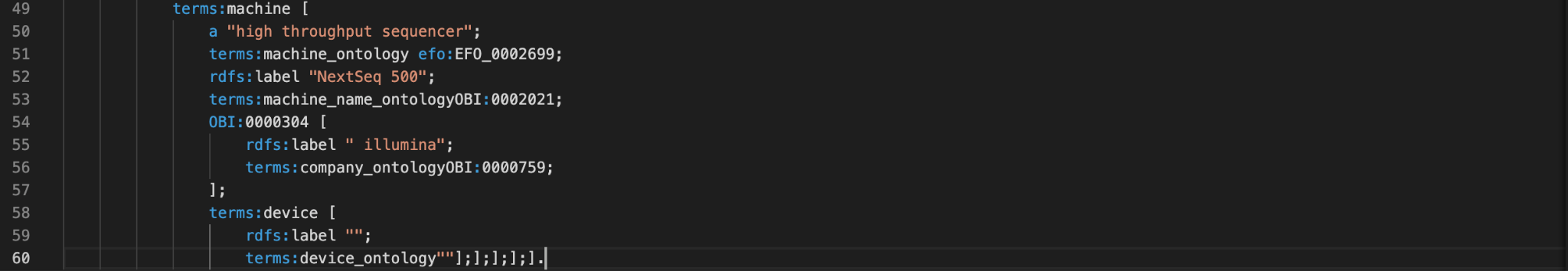
OBI:0000304 [の後、terms:company\_ontologyの前に改行を追加



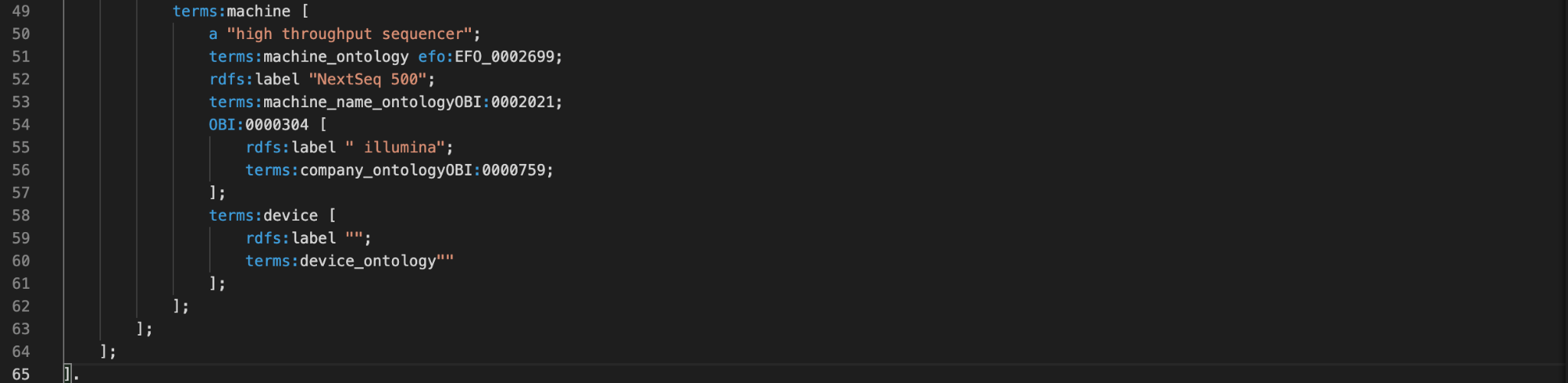
];terms:deviceの前に改行を入れて微調整



terms:deviceの前、rdfs:label、terms:device\_ontologyの前に改行を入れる



微調整



# <最終確認>

稀に文字コードか何かでvalidatorが通らないことがある。そういう時はスペースが半角ではなく全角の場合がある

誤　terms:data\_type　"Controlled-Access (Type I)";

正　terms:data\_type "Controlled-Access (Type I)";

このようなミスはrdfs:resource、terms:data\_type、 terms:disease\_idの欄で頻発しやすいので要注意。

# <完成版>

humandbs:hum0249

dcterms:title "Elucidation of pathological conditions and development of therapeutic methods using iPS cells derived from patients with hard tissue/connective tissue related diseases";

dcterms:contributor [

a foaf:Person;

foaf:name "Toshifumi Azuma";

foaf:firstName "Toshifumi";

foaf:lastName "Azuma";

];

dcterms:references [

dc:title "CAGE-seq analysis of osteoblast derived from cleidocranial dysplasia human induced pluripotent stem cells";

rdfs:resource JGA:JGAD000348;

dcterms:identifier "doi:10.1016/j.bone.2020.115582";

prism:doi "10.1016/j.bone.2020.115582";

prism:url <http://dx.doi.org/10.1016/j.bone.2020.115582>;

];

pav:version [

owl:versionInfo 1;

dcterms:modified "2020/9/1"^^xsd:date;

terms:dataset [

terms:dataset\_member [

terms:study\_id JGA\_STUDY:JGAS000248;

terms:data\_type "Controlled-Access (Type I)";

];

terms:technology [

rdfs:label "CAGE-seq";

terms:technology\_ontology iobc:200906098595164768;

terms:JGA\_ID JGA\_STUDY:JGAS000248;

terms:participant [

rdfs:label "hard tissue/connective tissue related diseases (ICD10: Q870): 3 cases, osteoblast differentiated from iPS cells established from oral mucosa fibroblast: 2 cases (2 samples),osteoblast differentiated from iPS cells established from oral mucosa fibroblast (mutation-corrected iPS cells by gene editing): 1 case (1 sample)";

terms:participant\_member [

terms:status "hard tissue/connective tissue related diseases";

terms:individuals "";

terms:sample\_number 3;

terms:disease\_id icd\_10:Q87.0;

];

terms:participant\_member [

terms:status "osteoblast differentiated from iPS cells established from oral mucosa fibroblast";

terms:individuals 2;

terms:sample\_number 2;

terms:disease\_id "";

];

terms:participant\_member [

terms:status "hard tissue/connective tissue related diseases";

terms:individuals 1;

terms:sample\_number 1;

terms:disease\_id "";

];

];

terms:machine [

a "high throughput sequencer";

terms:machine\_ontology efo:EFO\_0002699;

rdfs:label "NextSeq 500";

terms:machine\_name\_ontology OBI:0002021;

OBI:0000304 [

rdfs:label " illumina";

terms:company\_ontology OBI:0000759;

];

terms:device [

rdfs:label "";

terms:device\_ontology""

];

];

];

];

].

他参考：[NBDCヒトデータベースRDF化-浜口-0618.xlsx](https://drive.google.com/file/d/1rdFwlZTPE4QVnRDpRcdn_gjvHCT_M9XQ/view)

今までのトラブル対応

## (1)初期　~ttl文入力前作業~

Evernoteに作業の内容を整理しているのでそちらも参照

[NBDCヒトデータベースのRDF化作業 (第一弾)](https://www.evernote.com/shard/s701/client/snv?noteGuid=e16a82d7-25f5-d63d-9400-7810a6a52107&noteKey=c85db0c7aaecaed25d591e8a1a321cb2&sn=https%3A%2F%2Fwww.evernote.com%2Fshard%2Fs701%2Fsh%2Fe16a82d7-25f5-d63d-9400-7810a6a52107%2Fc85db0c7aaecaed25d591e8a1a321cb2&title=NBDC%25E3%2583%2592%25E3%2583%2588%25E3%2583%2587%25E3%2583%25BC%25E3%2582%25BF%25E3%2583%2599%25E3%2583%25BC%25E3%2582%25B9%25E3%2581%25AERDF%25E5%258C%2596%25E4%25BD%259C%25E6%25A5%25AD%2B%2528%25E7%25AC%25AC%25E4%25B8%2580%25E5%25BC%25BE%2529)

[NBDCヒトデータベースのRDF化作業 (第二弾)](https://www.evernote.com/shard/s701/client/snv?noteGuid=7835d9f8-6b9b-2619-eb3a-3adf63944d5a&noteKey=3d9d43a387d531b0c592d6dea8198967&sn=https%3A%2F%2Fwww.evernote.com%2Fshard%2Fs701%2Fsh%2F7835d9f8-6b9b-2619-eb3a-3adf63944d5a%2F3d9d43a387d531b0c592d6dea8198967&title=NBDC%25E3%2583%2592%25E3%2583%2588%25E3%2583%2587%25E3%2583%25BC%25E3%2582%25BF%25E3%2583%2599%25E3%2583%25BC%25E3%2582%25B9%25E3%2581%25AERDF%25E5%258C%2596%25E4%25BD%259C%25E6%25A5%25AD%2B%2528%25E7%25AC%25AC%25E4%25BA%258C%25E5%25BC%25BE%2529)

[NBDCヒトデータベースのRDF化作業 (第三弾)](https://www.evernote.com/shard/s701/client/snv?noteGuid=7ea9eb15-067c-63ee-edf3-9940eb67246d&noteKey=945c712ac9e082eaca78bbfe20af78dc&sn=https%3A%2F%2Fwww.evernote.com%2Fshard%2Fs701%2Fsh%2F7ea9eb15-067c-63ee-edf3-9940eb67246d%2F945c712ac9e082eaca78bbfe20af78dc&title=NBDC%25E3%2583%2592%25E3%2583%2588%25E3%2583%2587%25E3%2583%25BC%25E3%2582%25BF%25E3%2583%2599%25E3%2583%25BC%25E3%2582%25B9%25E3%2581%25AERDF%25E5%258C%2596%25E4%25BD%259C%25E6%25A5%25AD%2B%2528%25E7%25AC%25AC%25E4%25B8%2589%25E5%25BC%25BE%2529)

### <表を一瞥してわかりにくいところの列挙>

・NBDC Research IDのhum…と分子データIDのJGA…/DRA…の表記が並列で表示されて

いる

→NBDC Research ID下に別々の分子データIDが来るようにttl文を書く

・v1:YYYY/MM/XX　version番号と公開日が合体して表示

→version番号と公開日が対応するようなttl文が必要

・手法はNGS（RNA-seq、CAGE-seq）のような括弧内での並列表示あり

→NGSは消して中の表記を採用し、手法は並列表記するような形が必要あり

・機器のMRIスキャナーにはリンクが貼ってある

→リンク先の情報も抽出して書く必要あり

・データについて制限公開と非制限公開の並列表示あり+表記揺れあり

→表記を定めて、制限公開と非制限公開がどのデータと対応するかわかるようにする必要あり

・著者によってはグループ名が記載されている

→グループ名は今回は入れない方針

### <各カラムの要素数確認>~2021年1月時点~

データの種類(日本語版)

({'NGS（Exome）': 58, 'NGS（RNA-seq）': 45, 'NGS（Target Capture）': 28, 'SNP-chip': 27, 'NGS（WGS）': 21, 'NGS（ChIP-seq）': 9, 'メチル化アレイ': 8, 'NGS（WGBS）': 7, 'NGS（CAGE-seq）': 6, '臨床情報': 3, 'NGS（scRNA-seq）': 2, 'メタゲノム': 2, 'NGS（RRBS）': 2, 'マイクロアレイ': 2, 'NGS（small RNA-seq）': 2, '脳画像': 2, 'LC-MS/MS': 1, 'ダイレクトシークエンス': 1, 'NGS（Hi-C）': 1, 'eQTL': 1, '病理スライドガラス画像': 1, 'NGS（PBAT-seq）': 1, 'NGS（ATAC-seq）': 1, 'Sanger-seq': 1, 'NGS（miRNA-seq）': 1, 'NGS（Amplicon-seq）': 1, ' NGS（ChIP-seq）': 1, 'アレイCGH': 1, 'NGS（PCR amplicon Seq）': 1, '\*サンガー': 1, 'TaqMan': 1, 'qRT-PCR': 1, 'COBRA': 1, '治療情報': 1, 'NGS（Exome Target Capture）': 1, 'NGS（BS-seq）': 1, 'NGS（PBAT）': 1})

データの種類(英語版)

({'NGS (Exome)': 59, 'NGS (RNA-seq)': 45, 'NGS (Target Capture)': 29, 'SNP-chip': 24, 'NGS (WGS)': 22, 'NGS (ChIP-seq)': 10, 'Methylation array': 8, 'NGS (WGBS)': 7, 'NGS (CAGE-seq)': 6, 'Control Set': 3, 'NGS (scRNA-seq)': 2, 'Direct sequencing': 2, 'Metagenome': 2, 'NGS (RRBS)': 2, 'NGS (small RNA-seq)': 2, 'Neuroimage': 2, 'Clinical Info': 2, 'Case-Control': 2, 'LC-MS/MS': 1, 'NGS (Hi-C)': 1, 'eQTL': 1, 'Immunohistochemistry': 1, 'NGS (PBAT-seq)': 1, 'NGS (ATAC-seq)': 1, 'SNP array': 1, ' SNP-chip': 1, 'Sanger-seq': 1, 'Microarray': 1, 'NGS (miRNA-seq)': 1, 'NGS (Amplicon-seq)': 1, 'NGS (SNP-chip)': 1, 'Array CGH': 1, 'NGS (PCR amplicon Seq)': 1, 'TaqMan': 1, 'qRT-PCR': 1, 'COBRA': 1, 'cDNA microarray': 1, 'Clinical data': 1, 'Treatment data': 1, 'NGS (BS-seq)': 1, 'Platform comparison': 1, 'CNV Analysis': 1, 'Tumor vs. Matched-Normal': 1, 'Cohort': 1, 'NGS (PBAT)': 1, 'Single Patient': 1})

研究方法の要素数

{'配列決定': 98, '発現': 49, 'メチル化': 20, 'ゲノムワイド関連': 18, 'ヒストン修飾': 10, 'ゲノムワイド': 7, '転写開始点同定': 5, 'クロマチン構造': 2, 'eQTL': 2, 'メタゲノム': 2, 'CNV': 2, '健常者遺伝子型': 2, '脳画像ベースの病態解明': 2, 'HLA領域': 2, '健常者内頻度': 2, '質量分析': 1, '病理画像': 1, 'キャプチャ方法の性能比較': 1, 'HLAアリルおよび領域内遺伝子型決定': 1, 'HBV組込部位': 1, 'メチル化シトシン同定': 1}

Platformの要素数

({'Illumina (HiSeq 2000)': 38, 'Illumina (HiSeq 2500)': 37, 'Illumina (HiSeq 2000/2500)': 12, 'Illumina (MiSeq)': 11, 'Illumina (NovaSeq 6000)': 7, 'Illumina (Genome Analyzer IIx)': 7, 'Thermo Fisher Scientific (Ion Proton)': 6, 'Illumina (HiSeq 1500/2500)': 5, 'Affymetrix (Axiom ASI)': 5, 'Illumina (HiSeq X Ten)': 4, 'Illumina (NextSeq 500)': 3, 'Illumina (HiSeq 1500)': 3, 'Nanopore (MinION)': 3, 'Illumina (HiSeq 2500/4000)': 3, 'Thermo Fisher Scientific (Ion S5)': 3, 'Affymetrix (SNP Array 6.0)': 3, 'Affymetrix (Japonica Array v1)': 2, 'Thermo Fisher Scientific (Ion Torrent)': 2, 'Illumina (HiSeq)': 2, 'Illumina (IInfinium OmniExpressExome BeadChips)': 2, 'Nanopore (PromethION)': 2, '10x Genomics (Chromium Controller)': 2, 'Illumina (NextSeq)': 2, 'Illumina (HumanExome BeadChip)': 2, 'Illumina (HumanCoreExome BeadChip)': 2, 'Illumina (Human Methylation 450K BeadChip)': 2, 'Illumina (Infinium MethylationEPIC)': 2, 'Illumina (HumanOmniExpress BeadChip)': 2, 'Affymetrix (Axiom Genome-Wide ASI 1 Array)': 1, 'PacBio (RS II)': 1, 'Bionano (Irys)': 1, ' Bionano (Saphyr)': 1, 'Illumina (HiSeq 2500/3000)': 1, 'Sciex (TripleTOF 5600)': 1, 'Thermo Fisher Scientific (GeneChip® Scanner 3000 7G)': 1, 'Applied Biosystems (Veriti 96-Well Thermal Cycler and 3730XL DNA Analyzer)': 1, 'Illumina (HiSeq series)': 1, 'Illumina (HiSeq 3000)': 1, 'Illumina (Infinium Asian Screening Array)': 1, 'Illumina (HumanOmniExpress-24 BeadChip)': 1, 'Illumina (HumanOmniExpressExome-8 BeadChip)': 1, 'Illumina (HiSeq 1500/4000)': 1, 'Dako (Autostainer Link 48)': 1, 'PacBio (Sequel)': 1, 'Illumina (HumanOmniExpressExome-8)': 1, 'Illumina (660W-Quad)': 1, 'Illumina (Omni2.5-8)': 1, 'Illumina (OmniExpress-12)': 1, 'Illumina (OmniExpress-24)': 1, 'Illumina (HumanOmniExpressExome)': 1, 'Illumina (HumanOmniExpress)': 1, 'Illumina (Infinium OmniExpressExome-8)': 1, 'Illumina (HiSeq 2000/2500/X Five)': 1, 'Illumina (Infinium MethylationEPIC BeadChip)': 1, 'Illumina (Infinium HumanMethylation450 BeadChip)': 1, 'Affymetrix (GeneChip Human Mapping 250K Nsp)': 1, 'Affymetrix (CytoScan HD)': 1, 'Affymetrix (Japonica Array v1/v2)': 1, 'Illumina (Infinium Omni2.5-8 v1.4 BeadChip)': 1, 'Illumina (Human610-Quad BeadChip)': 1, 'Illumina (HumanOmni2.5-Quad BeadChip)': 1, 'Thermo Fisher Scientific (ABI) 3500 Genetic Analyzer': 1, 'Agilent Technologies (SurePrint G3 Human GE v2 8x60K Microarray)': 1, 'Agilent Technologies (SureScan G4900DA)': 1, 'Illumina (Infinium HumanMethylation450)': 1, 'Illumina (Infinium Human MethylationEPIC BeadChip)': 1, 'Illumina (HumanOmni2.5-8 BeadChip)': 1, 'Illumina (HiSeq 4000/X Ten)': 1, 'Affymetrix (Axiom ASI/6.0)': 1, 'GA 3130': 1, 'LightCycler 480II': 1, 'Whole Human Genome Microarray 4x44K': 1, 'Illumina (HiSeq 2500/X)': 1, 'MRI Scanners (15 Models)': 1, 'PET Scanners (16 Models)': 1, 'Illumina (HiSeq 1500)': 1, 'Affymetrix (GeneChip Human Mapping 250k Nspl)': 1, 'Affymetrix (CytoScan HD arrays)': 1, 'Philips (Achieva)': 1, 'Affymetrix (Human Mapping 250K Nsp Array)': 1, 'Illumina (ImmunoChip)': 1, 'Illumina (HumanHap550v3 Genotyping BeadChip)': 1, 'Illumina (Human610-Quad': 1, 'Illumina (HumanHap550v3)': 1, 'Illumina (HumanOmniExpress-12)': 1, 'Illumina (HumanExome)': 1, 'Illumina (OmniExpressExome)': 1, 'Perlegen Sciences (high-density oligonucleotide arrays)': 1, 'Hologic Japan (Invader)': 1, 'Illumina (HumanOmni2.5-8)': 1, 'Illumina (Human610-Quad)': 1, 'Illumina (HumanOmni2.5-4/2.5-8/2.5S-8)': 1, 'Illumina (HumanExome-12)': 1, 'Illumina (HumanCoreExome-12)': 1, 'Roche GS FLX Titanium': 1, 'Applied BioSystems (AB5500 GA)': 1, 'Illumina (HumanMethylation450 BeadChip)': 1, 'Life technoligies (Ion PGM/IonProton)': 1})

→chip名やkit名と機器名が混ざった表記になっているので分ける必要がある

### <日本語版と英語版の比較>

・英語版には、「研究方法」のカラムがないことが判明

・英語版には、hum番号[hum0003.v1](https://humandbs.biosciencedbc.jp/hum0003-v1)自体が記載されていないことが判明

→とりあえず保留

### <Topページと詳細ページの比較>

ex)hum0201.v3

topページのPlattform

(HiSeq 2500/4000, NovaSeq 6000)

vs

詳細ページのPlattform

(llumina [HiSeq 2500/4000], Illumina [HiSeq X ten], Illumina [HiSeq2500/NovaSeq6000], Illumina [HiSeq X ten])

→topページはversion更新がされていない + 各データで使った機器詳細表記できていないので詳細ページを見てttl文を書く必要あり

### <機器名のリスト化>

Illumina

'Illumina (HiSeq 1500)',

'Illumina (660W-Quad)',

'Illumina (Genome Analyzer IIx)',

'Illumina (HiSeq 1500)',

'Illumina (HiSeq 1500/2500)',

'Illumina (HiSeq 1500/4000)',

'Illumina (HiSeq 2000)',

'Illumina (HiSeq 2000/2500)',

'Illumina (HiSeq 2000/2500/X Five)',

'Illumina (HiSeq 2500)',

'Illumina (HiSeq 2500/3000)',

'Illumina (HiSeq 2500/4000)',

'Illumina (HiSeq 2500/X)',

'Illumina (HiSeq 3000)',

'Illumina (HiSeq 4000/X Ten)',

'Illumina (HiSeq X Ten)',

'Illumina (HiSeq series)',

'Illumina (HiSeq)',

'Illumina (Human Methylation 450K BeadChip)',

'Illumina (Human610-Quad',

'Illumina (Human610-Quad BeadChip)',

'Illumina (Human610-Quad)',

'Illumina (HumanCoreExome BeadChip)',

'Illumina (HumanCoreExome-12)',

'Illumina (HumanExome BeadChip)',

'Illumina (HumanExome)',

'Illumina (HumanExome-12)',

'Illumina (HumanHap550v3 Genotyping BeadChip)',

'Illumina (HumanHap550v3)',

'Illumina (HumanMethylation450 BeadChip)',

'Illumina (HumanOmni2.5-4/2.5-8/2.5S-8)',

'Illumina (HumanOmni2.5-8 BeadChip)',

'Illumina (HumanOmni2.5-8)',

'Illumina (HumanOmni2.5-Quad BeadChip)',

'Illumina (HumanOmniExpress BeadChip)',

'Illumina (HumanOmniExpress)',

'Illumina (HumanOmniExpress-12)',

'Illumina (HumanOmniExpress-24 BeadChip)',

'Illumina (HumanOmniExpressExome)',

'Illumina (HumanOmniExpressExome-8 BeadChip)',

'Illumina (HumanOmniExpressExome-8)',

'Illumina (IInfinium OmniExpressExome BeadChips)',

'Illumina (ImmunoChip)',

'Illumina (Infinium Asian Screening Array)',

'Illumina (Infinium Human MethylationEPIC BeadChip)',

'Illumina (Infinium HumanMethylation450 BeadChip)',

'Illumina (Infinium HumanMethylation450)',

'Illumina (Infinium MethylationEPIC BeadChip)',

'Illumina (Infinium MethylationEPIC)',

'Illumina (Infinium Omni2.5-8 v1.4 BeadChip)',

'Illumina (Infinium OmniExpressExome-8)',

'Illumina (MiSeq)',

'Illumina (NextSeq 500)',

'Illumina (NextSeq)',

'Illumina (NovaSeq 6000)',

'Illumina (Omni2.5-8)',

'Illumina (OmniExpress-12)',

'Illumina (OmniExpress-24)',

'Illumina (OmniExpressExome)'

Affymetrix

'Affymetrix (Axiom ASI)',

'Affymetrix (Axiom ASI/6.0)',

'Affymetrix (Axiom Genome-Wide ASI 1 Array)',

'Affymetrix (CytoScan HD arrays)',

'Affymetrix (CytoScan HD)',

'Affymetrix (GeneChip Human Mapping 250K Nsp)',

'Affymetrix (GeneChip Human Mapping 250k Nspl)',

'Affymetrix (Human Mapping 250K Nsp Array)',

'Affymetrix (Japonica Array v1)',

'Affymetrix (Japonica Array v1/v2)',

'Affymetrix (SNP Array 6.0)'

Thermo Fisher Scientific

'Thermo Fisher Scientific (ABI) 3500 Genetic Analyzer',

'Thermo Fisher Scientific (GeneChip® Scanner 3000 7G)',

'Thermo Fisher Scientific (Ion Proton)',

'Thermo Fisher Scientific (Ion S5)',

'Thermo Fisher Scientific (Ion Torrent)'

'10x Genomics (Chromium Controller)',

'Agilent Technologies (SurePrint G3 Human GE v2 8x60K Microarray)',

'Agilent Technologies (SureScan G4900DA)',

'Applied BioSystems (AB5500 GA)',

'Applied Biosystems (Veriti 96-Well Thermal Cycler and 3730XL DNA Analyzer)',

'Bionano (Irys)',

'Bionano (Saphyr)',

'Dako (Autostainer Link 48)',

'GA 3130',

'Hologic Japan (Invader)',

'Life technoligies (Ion PGM/IonProton)',

'LightCycler 480II',

'MRI Scanners (15 Models)',

'Nanopore (MinION)',

'Nanopore (PromethION)',

'PET Scanners (16 Models)',

'PacBio (RS II)',

'PacBio (Sequel)',

'Perlegen Sciences (high-density oligonucleotide arrays)',

'Philips (Achieva)',

'Roche GS FLX Titanium',

'Sciex (TripleTOF 5600)',

'Whole Human Genome Microarray 4x44K'

→各機器カッコの中身を出して書く必要あり

　HiSeq X TenとHiSeq X Fiveは「HiSeq X」に統一の方針

NextseqはNextseq 500に統一の方針

### <手法のOntologyシノニマス探索>

例えば、RNA-seqは以下のようなsynonymousがある

・MicroRNA sequencing

・RNA sequencing

・Small RNA-Seq

・Small RNA sequencing

・WTSS

・Transcriptome profiling

・Whole transcriptome shotgun sequencing

・RNA-Seq analysis

・Small-Seq

・miRNA-seq

→Ontologyでは同じものとして扱うように表記する

### <Topページと詳細ページの比較>

・患者情報がTopページと詳細ページで違った

[hum0243-v1](https://humandbs.biosciencedbc.jp/en/hum0243-v1)

誤　臍帯血：24名　急性骨髄性白血病：116症例

正　臍帯血：24名　急性骨髄性白血病：114症例

hum0122-v1

誤　市販のヒト培養細胞、およびそれから誘導したiPS細胞株：計27検体（細胞株）

正　市販のヒト培養細胞、およびそれから誘導したiPS細胞株：計28検体（細胞株)

## (2)中期　~csv入力作業~

初期時に判明したカラムの要素全てをcsvに分けて書く作業を行った

CSV to RDF convert 手順　の欄を参照のこと

### <Turtle Validator上でttl文確認作業中に文字コードのトラブル>

以下のサイトを参考にして修正した

<https://qiita.com/pugiemonn/items/106749351991037fb606>

<https://lab.unicast.ne.jp/2013/10/22/fix-c2a0-problem/>

## (3)後期　~ttl文入力作業~

### <ttl文のRDF format見直し>

terms:technologyとterms:machine以下を並列で書いていたが、

terms:technology以下にterms:machineを書く方が見やすい上に構造的に妥当と気づき、

今まで書いたRDF formatの修正を行った 　かなり骨が折れる作業だった

### <患者情報の追加・修正>

各hum番号の患者情報は各患者情報が総合して書かれているケースが多く、分割し、さらに各患者に対応するICD10情報や患者数、データ等の探索が必要。

特に画像情報がmainの研究(例：<https://humandbs.biosciencedbc.jp/hum0250-v1>)は

通常のformatとかなり異なっているので苦労した。

### <virtuosoを使ってのttl文の構造確認>

Server上のVirtuosoを起動してEndopoint　browserで構造を確認した

以下のVirtuosoの使用方法について記載しておく(三橋さんの説明をそのまま記載)

・起動停止の方法

hamaguchi@dgx1:/raid/work/humandbs/virtuoso-docker$ pwd

/raid/work/humandbs/virtuoso-docker

hamaguchi@dgx1:/raid/work/humandbs/virtuoso-docker$ docker-compose up -d

Creating network "virtuoso-docker\_default" with the default driver

Creating virtuoso-docker\_sparqlist\_1 ... done

Creating virtuoso-docker\_virtuoso\_1 ... done

Creating virtuoso-docker\_nginx\_1 ... done

hamaguchi@dgx1:/raid/work/humandbs/virtuoso-docker$ docker-compose ps

Name Command State Ports

--------------------------------------------------------------------------------------------

virtuoso-docker\_nginx\_1 /docker-entrypoint.sh ngin ... Up 0.0.0.0:10702->80/tcp

virtuoso-docker\_sparqlist\_1 docker-entrypoint.sh /bin/ ... Up

virtuoso-docker\_virtuoso\_1 /bin/bash /virtuoso.sh Up 1111/tcp, 8890/tcp

hamaguchi@dgx1:/raid/work/humandbs/virtuoso-docker$ docker-compose down

Stopping virtuoso-docker\_nginx\_1 ... done

Stopping virtuoso-docker\_sparqlist\_1 ... done

Stopping virtuoso-docker\_virtuoso\_1 ... done

Removing virtuoso-docker\_nginx\_1 ... done

Removing virtuoso-docker\_sparqlist\_1 ... done

Removing virtuoso-docker\_virtuoso\_1 ... done

Removing network virtuoso-docker\_default

hamaguchi@dgx1:/raid/work/humandbs/virtuoso-docker$ docker-compose up -d

Creating network "virtuoso-docker\_default" with the default driver

Creating virtuoso-docker\_sparqlist\_1 ... done

Creating virtuoso-docker\_virtuoso\_1 ... done

Creating virtuoso-docker\_nginx\_1 ... done

hamaguchi@dgx1:/raid/work/humandbs/virtuoso-docker$

・ロードの方法

virtuoso-dockerをホームディレクトリとします。

* bin/load.shをのadd\_load\_listに<<RDFファイルディレクトリ>> <<ファイル名>> <<グラフ名>>を追加します。

#add\_load\_list ${LOAD\_DATA\_BASE}/samples/mesh '\*.nt.gz' '<http://humandbs.biosciencedbc.jp/mesh>'

#add\_load\_list ${LOAD\_DATA\_BASE}/samples/mondo '\*.owl' '<http://humandbs.biosciencedbc.jp/mondo>'

#add\_load\_list ${LOAD\_DATA\_BASE} 'humandbs\_0520.ttl' '[http://humandbs.biosciencedbc.jp](http://humandbs.biosciencedbc.jp/)'

add\_load\_list ${LOAD\_DATA\_BASE}/samples 'humandbs\_0524.ttl' '<http://humandbs.biosciencedbc.jp/contributor>'

2. <<RDFファイルディレクトリ>> は、load以下です。適当な場所にファイルを置いてください。

3. bin/exec\_load.rdf.shを実行してロードします。

4. logs/exec\_load\_rdf\*.logにログが出力されます。Errorも出力されます。

5. Errorを修正してもErrorが取れない時は、以下の部分を実行して記録をクリアしてください。

## 特定のgraphのロードファイルの記録を削除

## <https://wiki.lifesciencedb.jp/mw/SPARQLthon19/TripleLoad>

#${ISQL} -H ${HOST} -S ${PORT} -U ${USER} -P ${PASSWORD} VERBOSE=OFF BANNER=OFF EXEC="DELETE FROM DB.DBA.LOAD\_LII

ST WHERE ll\_graph = '<http://humandbs.biosciencedbc.jp/contributor>';"

・削除の方法

bin/empty\_virtuoso.shを実行すると、データベースを初期化します。logs/empty\_virtuoso\*.logにログが出力されます。

特定のグラフだけを削除する場合は、drop\_graph.shのGRAPH=“<<http://humandbs.biosciencedbc.jp/contributor>>“を削除したいグラフ名に変更して実行してください。

今回はデータ量が少ないので、empty\_virtuoso.shして全部ロードしてもすぐ終わると思います。

内容はエンドポイントから確認できます。

<https://test64.biosciencedbc.jp/sparqlist/>

グラフの一覧は以下のSPARQListを実行しても確認できます。

<https://test64.biosciencedbc.jp/sparqlist/show_graphs>

やり残したこと・できなかったこと

### <更新できていないhumNo.>

2021/12/6以降のデータ(hum0317および120.v3以降)

### <hum0181について>

[https://humandbs.biosciencedbc.jp/en/hum0181-v1-st1](https://humandbs.biosciencedbc.jp/en/hum0181-v1-st1%E3%82%92%E5%8F%82%E7%85%A7%E3%81%AE%E3%81%93%E3%81%A8)を参照のこと。

各Caseごとのdataset名(大量)にあり、期間内では対応できなかった

### <患者情報>

性別や国籍等の情報は記載したが、他の患者情報(例えば年齢、幼年か少年か大人か老年かなど)については用意していないので、今後そのような情報についてはOntolohyを用意する必要がある。

### <提供者情報>

修正するに当たって参考になるRDF(<https://support.nii.ac.jp/en/cia/api/a_rdf_auth>)

例えば、hum199だと

研究代表者： 藤村 哲也

所 属 機 関： 東京大学医学部附属病院 泌尿器科

プロジェクト/研究グループ名： 次世代シーケンサーを用いた泌尿器悪性腫瘍における治療標的分子の同定

**科研費/助成金（Research Project Number）：**

| 科研費・助成金名 | タイトル | 研究課題番号 |
| --- | --- | --- |
| 科学研究費助成事業 基盤研究（B） | 前立腺癌のホルモン療法抵抗性獲得に至るエピゲノム調節機構の統合的解明と臨床応用 | 17H04334 |
| 科学研究費助成事業 基盤研究（C） | エストロゲン・アンドロゲンシグナルによる転移性前立腺癌のプレシジョンメディシン | 18K09128 |
| 科学研究費助成事業 基盤研究（C） | Genetic scoreを用いた前立腺癌治療の個別化 | 15K10576 |

と記載されている。

これらの項目をRDF化するのであれば、

dcterms:contributor [

a foaf:Person;

foaf:name "";

foaf:firstName "";

foaf:lastName ""

];

の欄を修正する形になる。ただ、直す際の注意点が多々ある。

まず所属機関だが、大学・大学院/学部・研究科・研究所/部署のような順に書き下していく必要があるかもしれない。その場合はそれぞれのランクごとにOntologyを用意する必要がある。

次に、プロジェクト/研究グループ名は日本語表記のみのものが多く、例えばhum0199のプロジェクト/研究グループ名を探してみると[リンク集 | 東大泌尿器科学教室](https://square.umin.ac.jp/UT-urology/link.html)　内の[次世代シーケンサーを用いた泌尿器悪性腫瘍における治療標的分子の同定](https://square.umin.ac.jp/UT-urology/pdf/doc14.pdf)にしか検索でヒットしなかった。Ontology等で統一できない項目なので、rdf:label等で、rdf:resourceやprism:urlやrdfs:seeAlsoなどを使ってリンクを教えてあげるのがbetterかもしれない。

案としては、例えばhum0199の場合だと以下のような修正案が考えられる。

dcterms:contributor [

a foaf:Person;

foaf:name "Tetsuya Fujimura";

foaf:firstName "Tetsuya";

foaf:lastName "Fujimura"

foaf:interest [

rdf:label "Identification of therapeutic target molecules for urologic cancers by using next generation sequencer";

rdf:resource <https://square.umin.ac.jp/UT-urology/pdf/doc14.pdf>;

];

また、科研費/助成金についてはKAKENHIの研究課題番号であれば

@prefix:はKAKENHI <https://kaken.nii.ac.jp/ja/grant/KAKENHI-PROJECT->.

を用意して番号と対応するタイトルと助成金名を当てていく作業が必要となる。

助成金名について例えば「科学研究費助成事業 基盤研究（B）」のオントロジーはないのでラベルのみでの対応になると思われる。

### <分子データ>

項目としては以下である

・対象領域（Target Captureの場合）

・ライブラリソース

・断片化の方法

・ライブラリ構築方法

・リード長（除：バーコード、アダプタ、プライマー、リンカー）

・総データ量

・コメント（利用にあたっての制限事項）

情報として必要かと言うと点で保留にしていた。

RDF化する場合は機械可読性にたける表現にする必要がある。

### <TogoDB上でのcsv→ttl変換の試み>