

La gestion de (méta)données terminologiques «FAIR»: le répertoire de catégories de données de la ressource TriMED

Federica Vezzani

Département d'études linguistiques et littéraires (DiSLL)
Via Elisabetta Vendramini, 13, 35137, Padoue - Italie
Université de Padoue
federica.vezzani@unipd.it
<http://www.dei.unipd.it/~vezzanif/>

Résumé. Cette étude est consacrée à la description d'un flux de travail pour la gestion optimale de données et métadonnées terminologiques figurant dans les ressources terminologiques. Le paradigme de la «Terminologie FAIR» sur lequel nous sommes basées vise à fournir des lignes directrices pour la mise à disposition de données de la recherche terminologique trouvables, accessibles, interopérables et réutilisables. En particulier, nous nous concentrons sur la description d'un nouveau répertoire de catégories de données terminologiques - conçues comme des classes d'information - spécifiquement implémenté à l'appui de la ressource terminologique multilingue et polyvalente TriMED pour le domaine médical. Le répertoire est structuré selon les lignes directrices fournies dans la norme ISO 12620: 2019 et représente l'une des étapes vers la «FAIRification» des données terminologiques contenues dans TriMED.

1. Introduction

L'implémentation et la mise à disposition d'une base de données terminologique nécessitent de nombreux efforts de conception préliminaire afin d'organiser et de gérer de manière optimale les (méta)données terminologiques.

«Gérer les données, c'est s'assurer que celles-ci sont correctement sélectionnées, décrites, préservées et rendues accessibles pour un traitement

et/ou une réutilisation, et ce, bien au-delà du projet de recherche qui les a fait naître et les a exploitées au premier chef».

C'est ainsi que les auteurs Calderan et Millet (2015) définissent l'ensemble de tâches relevant de l'activité de «curation de données» (de l'anglais *data curation*). Cette notion identifie donc, généralement, l'ensemble de bonnes pratiques pour l'organisation optimale des données de recherche (Palmer *et al.* 2013), dont la responsabilité incombe inévitablement au chercheur qui produit ces données (McLure *et al.* 2014, Corti *et al.* 2019). L'ampleur et l'importance de la curation des données ont conduit à l'émergence d'une littérature et de normes détaillées régissant les actions requises (De Matos *et al.* 2004, Eaker 2016, Erkimbaev *et al.* 2019). À cet égard, le respect de ces exigences peut garantir non seulement la sécurité, mais également un enrichissement continu de la valeur qualitative des données scientifiques. À ce propos, un ensemble de lignes directrices a été publié par Wilkinson *et al.* (2016) dans le cadre de la plateforme européenne, librement accessible en ligne, *European Open Science Cloud* (EOSC)¹, afin de promouvoir la «FAIRness» des données de recherche. Ces lignes directrices soulignent la nécessité de données trouvables, accessibles, interopérables et réutilisables (*Findable, Accessible, Interoperable* et *Reusable*)². Toutes les composantes du processus de recherche devraient bénéficier donc de l'application de ces principes, afin d'en garantir la transparence, la reproductibilité et la réutilisabilité.

La bonne gestion des données n'est pas un objectif en soi, mais plutôt le principal moyen menant à la découverte et à l'innovation des connaissances, ainsi qu'à l'intégration et à la réutilisation des données par la communauté scientifique. Dans le cadre de l'activité terminographique, l'Organisation internationale de normalisation (ISO)³ et, en particulier, le Comité technique ISO/TC 37 (Langage et terminologie)⁴ fournissent des normes spécifiques pour la conception et l'implémentation des ressources linguistiques et terminologiques structurellement homogènes. Toutefois, dans ce contexte, les données de la recherche sont encore loin d'être FAIR (Forkel *et al.* 2018). Les ressources langagières sont souvent encodées dans un format hétérogène et développées isolément les unes des autres (Cimiano *et al.* 2020), au risque de rendre leur découverte, leur réutilisation et leur intégration une tâche difficile et lourde. En ce sens, il faut noter les efforts menés par l'infrastructure

1 <https://www.eosc-portal.eu>

2 <https://www.go-fair.org/fair-principles/>

3 <https://www.iso.org/fr/home.html>

4 <https://www.iso.org/fr/committee/48104/x/catalogue/>

de recherche européenne CLARIN⁵ qui permet aux chercheurs en sciences humaines d'accéder aux ressources et technologies linguistiques disponibles au niveau européen et vise à fournir une architecture des données conforme aux principes FAIR (De Jong *et al.* 2018).

Notre proposition s'inscrit dans ce contexte et vise à la description d'une étude de cas portant sur l'organisation FAIR des (méta)données terminologiques de la ressource multilingue TriMED⁶ conçue pour le domaine d'application médical (Vezzani et Di Nunzio 2020b). La base de données terminologique a été implémentée dans l'esprit de la «terminologie FAIR» qui porte sur l'application des normes ISO les plus récentes sur la gestion des ressources terminologiques dans le but ultime de fournir à la communauté scientifique des données trouvables, accessibles, interopérables et réutilisables (Vezzani et Di Nunzio 2020a). En particulier, dans cette étude nous nous concentrerons sur le développement d'une ressource parallèle et complémentaire à TriMED, c'est-à-dire son répertoire de catégories de données⁷ qui a été spécifiquement conçu, en respectant la norme ISO 12620: 2019⁸, afin de documenter, d'harmoniser et de gérer de manière optimale les (méta)données terminologiques. L'organisation de l'article est la suivante. La section 2 est consacrée à la description de la notion de «catégories de données terminologiques». La section 3 illustre le cadre général dans lequel s'inscrit le projet de recherche TriMED et définit le paradigme de la «Terminologie FAIR» pour la structuration optimale des ressources terminologiques. En particulier, nous nous concentrons sur la description du répertoire de catégories de données spécifiquement conçu pour la ressource TriMED. Enfin, dans la section 4, nous présentons nos conclusions et perspectives de recherche à long terme.

2. Catégories de données terminologiques

Les données figurant dans les ressources linguistiques sont généralisées dans des classes appelées «catégories de données». Une catégorie de donnée est une classe d'informations étroitement liées d'un point de vue formel ou sémantique et fait partie d'un schéma de collecte de données ou d'annotation pour une ressource linguistique donnée (ISO 12620: 2019). Par exemple, /

5 <https://www.clarin.eu>

6 <https://purl.org/TriMED/>

7 La ressource est disponible au lien suivant : https://shiny.dei.unipd.it/TriMED/data_category_repository/

8 <https://www.iso.org/fr/standard/69550.html?browse=tc>.

définition/ et /partie du discours⁹ sont des catégories de données communes dans les ressources terminographiques et lexicographiques.

Toutefois, les variétés d'approches pour développer différents types de ressources linguistiques entraînent inévitablement des variations dans les dénominations et les définitions de catégories de données. En ce sens, le concept de cohérence des catégories de données est fondamental et se produit dans leur dénomination et dans les valeurs que ces catégories admettent (Warburton et Wright 2019). À des fins d'interopérabilité, par exemple, imaginons que nous voulions intégrer automatiquement les données de deux ressources terminologiques différentes qui présentent toutes deux la catégorie de donnée / définition/. Si le nom de la catégorie est différent (par exemple *définition* d'une part et *déf.* d'autre part), il serait difficile pour un logiciel d'interpréter les deux noms comme figurant la même catégorie et donc leur intégration ne serait pas automatique, mais demanderait un travail chronophage de nettoyage manuel (ou semi-automatique) des données.

Afin d'assurer la cohérence entre les catégories de données terminologiques représentées dans différentes ressources terminologiques (et pour que celles-ci «dialoguent» du point de vue structurel), un répertoire normalisé de catégories de données a été mis à la disposition avec la publication de la dernière version de la norme ISO-12620 de 2019. Le répertoire TermWeb, disponible sur le site DatCatInfo¹⁰, remplace l'ancienne ressource ISOcat développée et maintenue sous les auspices de l'ISO TC/37, avec l'Institut Max Planck de psycholinguistique de Nimègue, aux Pays-Bas, agissant en tant qu'autorité d'enregistrement (Kemps-Snijders *et al.* 2008, Broeder *et al.* 2014, Windhouwer et Schuurman 2014). Le nouveau répertoire TermWeb recueille une liste de toutes les catégories de données disponibles pour ceux qui souhaitent implémenter une ressource terminologique normalisée. Les catégories de données sont organisées en fiches descriptives et formelles (sous le nom de «spécifications») qui collectent toutes les informations nécessaires à leur documentation et à leur harmonisation, par exemple : nom, définition, exemples, commentaires, etc.

Par ailleurs, un aspect fondamental introduit par la norme ISO-12620 de 2019 est la possibilité pour tout implémenteur de concevoir son propre répertoire de catégories de donnée spécifique pour une ressource terminologique.

9 À des fins de cohérence avec les normes, nous utiliserons les barres obliques inversées lorsqu'indiquant une catégorie de donnée.

10 <http://datecatinfo.net>

Cette flexibilité de la norme est essentielle lorsqu'il est nécessaire de documenter des catégories de données qui n'existent pas dans le répertoire officiel TermWeb. De plus, les catégories de données ont des valeurs qui peuvent varier en fonction de la ressource conçue. La norme ISO-12620 de 2019 est complémentaire à la norme ISO 16642 : 2017¹¹ qui décrit le métamodèle structurel *Terminological Markup Framework* (TMF) à adopter pour la mise à disposition de ressources terminologiques interopérables. Le cadre de balisage terminologique TMF comprend un métamodèle structurel composé de trois entités principales et hiérarchiques (concept, langue et terme) et un ensemble de catégories de données pouvant être associées, selon une certaine liberté¹², à ces trois niveaux. En ce sens, la mise en œuvre parallèle d'un répertoire de catégories de données spécifiques pour une nouvelle ressource s'avère être une action indispensable non seulement pour la documentation correcte des (méta)données terminologiques, mais aussi pour désambiguïser leur signification et pour expliciter leurs contraintes en favorisant, de cette manière, le « dialogue structurel » entre les différentes ressources terminologiques et leur « FAIRness ».

3. Le projet TriMED

Avant de décrire l'objet de cette étude, à savoir l'élaboration d'un répertoire pour la documentation et la gestion optimale des catégories de données et des métadonnées terminologiques, nous présentons le cadre du projet de recherche dans lequel cette ressource s'inscrit. Le projet TriMED¹³ découle de la volonté de fournir une ressource multilingue et polyvalente pour le domaine médical, et structurée afin de respecter les besoins sous-jacents aux pratiques de la science ouverte (Vezzani et Di Nunzio 2019; Vezzani et Di Nunzio 2020a). Dans ce contexte, le projet a envisagé le développement du paradigme de la « Terminologie FAIR » proposant des lignes directrices pour la conception et l'implémentation optimale des ressources terminologiques (Vezzani 2020). Le paradigme vise à définir un flux de travail pour la mise à la disposition de (méta) données terminologiques trouvables, accessibles, interopérables et réutilisables et est basé sur l'adoption des dernières normes ISO TC/37 SC/3 pour la gestion terminologique :

11 <https://www.iso.org/fr/standard/56063.html?browse=tc>

12 Par exemple, selon la norme les éléments <term>, <termNoteGrp> et <termNote> peuvent uniquement apparaître au niveau du terme (<termSec>).

13 <https://purl.org/TriMED/>

- la norme ISO16642 : 2017¹⁴ qui définit le métamodèle structurel abstrait *Terminological Markup Framework* (TMF) pour la représentation des ressources terminologiques ;
- la norme ISO12620 : 2019¹⁵ qui définit les propriétés des catégories de données et leur documentation dans un répertoire ; et
- la norme ISO30042 : 2019¹⁶ qui définit le format de représentation *Term-Base eXchange* (TBX) des données terminologiques.

Sur la base des lignes directrices du paradigme proposé, afin qu'une ressource terminologique puisse être FAIR, elle devrait être conçue et implémentée en :

- suivant un modèle structurel interopérable (TMF) ;
- permettant l'accès aux données terminologiques via des protocoles de communication standard ;
- fournissant des (méta)données rigoureusement documentées, et par conséquent trouvables, à travers un répertoire de catégories de données ;
- garantissant la réutilisation des données grâce à l'application de formats pour l'échange terminologique (TBX).

Ces principes ont été adoptés pour le développement de la base de données multilingue (italien, français et anglais) TriMED qui vise à satisfaire les besoins d'information de différentes catégories d'utilisateurs (patients, traducteurs et médecins) en fournissant un modèle de fiche terminologique contenant jusqu'à 42 catégories de données terminologiques. En particulier, la ressource a été conçue afin de :

- aider les patients à comprendre correctement les informations médicales, compte tenu de l'aspect de variation diastratique de la terminologie ;
- soutenir le traducteur dans le processus de traduction spécialisée en fournissant un cadre sur le comportement syntaxique, sémantique et phraséologique du terme source et de son traduisant en langue cible ;
- fournir un point d'accès unique pour la consultation des professionnels de la santé aux autres terminologies, nomenclatures ou codes de classification internationaux généralement utilisés par des experts.

14 <https://www.iso.org/fr/standard/56063.html?browse=tc>.

15 <https://www.iso.org/fr/standard/69550.html?browse=tc>.

16 <https://www.iso.org/fr/standard/62510.html?browse=tc>.

La figure 1 représente un exemple d’affichage de la ressource pour la catégorie d’utilisateurs « Patient ».

The screenshot shows the 'TriMED Patient' interface. At the top, there is a blue header with the text 'TriMED Patient' and a hamburger menu icon. Below the header, there are two search filters: 'Langue' (Language) set to 'fr' and 'Recherche' (Search) set to 'Rougeole'. The main content area is divided into several sections: 'Définition' (Definition) with a paragraph describing the disease, 'Nom scientifique' (Scientific name) set to 'Rougeole', and 'Nom populaire' (Popular name) set to 'première maladie'.

FIG. 1 – Affichage « Patient » - TriMED

3.1. Répertoire de la ressource TriMED

Les exigences d’une structuration FAIR de la ressource TriMED nous ont menées à développer, en suivant la norme ISO-12620 de 2019, un répertoire parallèle et complémentaire contenant les spécifications de toutes les catégories de données affichées pour les utilisateurs. À notre connaissance, il s’agit du premier répertoire implémenté selon cette norme, dont l’application Web (déjà disponible en ligne)¹⁷ a été développée à l’aide du package *Shiny R* (Winston *et al.* 2018). L’utilisateur peut tout d’abord sélectionner la langue de travail et, ensuite, saisir la dénomination de la catégorie dans la boîte de recherche : le système filtrera automatiquement les mots et les caractères indiquant les options possibles. Une fois sélectionnée la catégorie de donnée, le système affichera en sortie automatique la spécification de la catégorie contenant toutes les informations requises par la norme afin d’en garantir la trouvabilité, l’accessibilité, l’interopérabilité et la réutilisabilité.

¹⁷ https://shiny.dei.unipd.it/TriMED/data_category_repository/

Data Category Repository - TriMED

Langue
fr

Catégorie de donnée
Collocation

Description XML

PID: <http://www.datcatinfo.net/datcat/DC-340>

Identifiant: collocation
Module: Trimed
Niveau (TMF): termSec
Classification: <termNote>
Typologie de contenu: chaîne de caractères
Valeur(s): NA

Description: Combinaison récurrente de mots caractérisée par la cohésion en ce que les composants de la collocation doivent coexister dans un énoncé ou une série d'énoncés, même s'ils ne doivent pas nécessairement maintenir une proximité immédiate les uns avec les autres.
Explication: Les collocations diffèrent des unités phraséologiques en ce que les composants de ces dernières doivent généralement apparaître dans une séquence fixe. Les combinaisons de mots récurrentes qui forment un terme complexe (par exemple, adjectif + nom, nom + nom, etc.) et qui représentent un concept unique ne sont pas des collocations.
Note: NA
Exemple: (Virus) attraper, contracter, inoculer, porter un ~; se protéger contre les ~. Un ~ se développe, se répand, se propage.

Download CSV

FIG. 2 – Spécification de la catégorie de donnée /collocation/.

La figure 2 ci-dessus montre une capture d'écran de la spécification de catégorie de donnée /collocation/. Pour chaque catégorie de données, les informations suivantes sont fournies :

1. Un identifiant unique et persistant (PID), c'est-à-dire une URL qui fournit l'accès Web direct à la spécification de la catégorie de donnée dans le répertoire en ligne.
2. Un identifiant mnémorique unique et stable de la catégorie de donnée qui ne doit pas inclure d'espaces entre les mots, car il est utilisé dans les environnements de codage comme élément ou comme valeur d'attribut.
3. Le module de catégories de données TBX auquel la catégorie se réfère¹⁸.

18 Le format d'implémentation choisi pour la ressource TriMED est le *TermBase Exchange* (TBX) conformément à la norme ISO 30042 : 2019 : <https://www.iso.org/fr/stand>

4. Le niveau du métamodèle TMF (concept, langue et terme) auquel la catégorie de donnée est associée dans notre ressource.
5. La typologie de contenu de la catégorie de donnée, c'est-à-dire les types d'informations que la catégorie de donnée permet par sa mise en œuvre, comme une «liste déroulante» ou une «chaîne de caractères».
6. L'ensemble de valeurs énumérées que la catégorie de donnée peut avoir si elle est implémentée en tant que «liste déroulante».
7. La définition de la catégorie de donnée.
8. D'autres explications et notes sur la catégorie de donnée.
9. Quelques exemples d'utilisation de la catégorie de donnée.
10. La traduction du nom canonique de la catégorie de donnée dans les autres langues de travail de la ressource.

Il faut noter que, afin d'assurer une certaine traçabilité, les informations fournies dans le répertoire de TriMED concernant les PID des catégories de données se réfèrent à l'URL correspondant sur le site DatCatInfo pour les catégories qui sont déjà documentées dans le répertoire TermWeb (voir figure 2). Il existe également un nombre restreint de catégories de données qui sont exclusives pour notre ressource et ne sont donc pas illustrées sur DatCatInfo. Ces catégories de données terminologiques comprennent : 1) /analyse sémique/, 2) /hyperonyme/, 3) /hyponyme/, 4) /sous-domaine/, 5) /code ICPC2/, 6) /code ICD10/, 7) /terme SNOMED CT/, 8) /terme MeSh/ et 9) /sphère conceptuelle/. Pour ces catégories, nous fournissons un PID qui correspond à l'URL de la page associée dans notre répertoire TriMED. Par exemple, pour la catégorie de donnée /analyse sémique/ le PID fourni est : <http://purl.org/trimed/dcr/dc/dc-1>. En outre, comme suggéré par la norme et pour répondre au besoin de réutilisation des données, l'utilisateur peut exporter les informations fournies dans le répertoire dans le format Comma-Separated Values (CSV) et eXtensible Markup Language (XML). Enfin, le répertoire de TriMED diffère du répertoire TermWeb en ce que, contrairement à ce dernier dans lequel seules des traductions du nom canonique des catégories de données sont proposées, notre application fournit à l'utilisateur un système d'affichage multilingue complet. En ce sens, des informations telles que «description», «explication»,

dard/62510.html?browse=tc. Chaque catégorie de données appartient donc à un module TBX (public ou privé). Pour plus d'informations sur cet aspect, voir Vezzani et Di Nunzio (2020a).

«notes» et «exemples» sont adaptées en fonction de la langue de consultation de l'utilisateur. Ce type d'affichage multilingue se reflète également dans la génération automatique de fichiers d'exportation afin que l'utilisateur puisse télécharger et réutiliser une spécification dans un format lisible par la machine et adapté à sa langue de travail.

4. Conclusions et perspectives

La gestion FAIR des (méta)données de la recherche terminologique demande un énorme travail de conception préliminaire afin d'organiser des informations structurellement homogènes. Ce type d'approche nécessite également une certaine réflexion sur l'activité terminographique: le chercheur se trouve donc à se pencher non seulement sur l'élaboration des données terminologiques pertinentes pour le domaine d'étude, mais aussi sur leur structuration optimale afin de permettre leur partage et leur réutilisation par la communauté scientifique. Dans cette étude, nous avons décrit le développement du répertoire de catégories de données de TriMED, en tant qu'une première étape vers une «terminologie FAIR». Les étapes présentées dans cet article couvrent la plupart des éléments clés du processus de FAIRification des données terminologiques. En perspective, nous entendons également définir les moyens de transformation de notre modèle sémantique en conformité avec le mouvement *Linguistic Linked Open Data* (LLOD). En ce sens, notre proposition contient un modèle conceptuel - schéma entité-association (Chen 1976) - qui peut être utilisé pour obtenir le graphe associé et transformer les données obtenues à partir de XML en RDF (*Resource Description Framework*) en tant que modèle de référence de données liées (Chiarcos *et al.* 2012).

Enfin, nous travaillons également à l'adoption de la même méthodologie d'organisation des données dans le cadre du projet européen «Terminologie sans frontières» promu par l'Unité de Coordination de la Terminologie (TermCoord) du Parlement européen. Le projet implique la collaboration de nombreux partenaires et requiert une action nécessaire de documentation et d'harmonisation des données dans l'esprit de la science ouverte. Dans ce cadre, nous avons mis en place et testons l'ergonomie de la nouvelle application web «FAIRterm¹⁹» (Vezzani 2021) comme outil de compilation de fiches terminologiques multilingues dans le cadre du projet européen. Cet outil vise donc à mettre à disposition des terminologues travaillant dans toute l'Europe une plateforme en accès libre permettant l'échange et le partage de données.

19 <https://purl.org/fairterm>.

Références

- Broeder, Daan, Ineke Schuurman, and Menzo Windhouwer. 2014. “Experiences with the ISOcat data category registry.” In *LREC 2014: 9th International Conference on Language Resources and Evaluation*, pp. 4565-4568.
- Calderan, L. et J. Millet. 2015. *BIG DATA : nouvelles partitions de l'information : Actes du séminaire IST Inria, octobre 2014*. Louvain-la-Neuve : De Boeck Supérieur.
- Winston, C., J. Cheng, J. J. Allaire, Y. Xie, et J. McPherson. 2018. “Shiny: Web Application Framework for R. R package version 1.1. 0.”
- Chen, Peter Pin-Shan. 1976. “The entity-relationship model—toward a unified view of data.” *ACM transactions on database systems (TODS)* 1, no. 1 : 9-36.
- Chiarcos, Christian, Sebastian Nordhoff, et Sebastian Hellmann. 2012. *Linked Data in Linguistics*. Heidelberg : Springer.
- Cimiano, Philipp, Christian Chiarcos, John P. McCrae, et Jorge Gracia. 2020. *Linguistic Linked Data*. Cham, Switzerland : Springer International Publishing.
- Corti, Louise, Veerle Van den Eynden, Libby Bishop, et Matthew Woollard. 2019. *Managing and sharing research data : a guide to good practice*. Los Angeles : SAGE Publications Limited.
- De Matos, David Martins, Ricardo Daniel Santos Faro Marques Ribeiro, et Nuno J. Mamede. 2004. “Rethinking reusable resources”. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).
- De Jong, F. M. G., Bente Maegaard, Koenraad De Smedt, Darja Fišer, et Dieter Van Uytvanck. 2018. “CLARIN: towards FAIR and responsible data science using language resources.” In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 3259-3264.
- Eaker, Chris. 2016. “What could possibly go wrong? the impact of poor data management”. In *The Medical Library Association's Guide to Data Management for Librarians*. Lanham, Maryland : Rowman and Littlefield Publishing Group.
- Erkimbaev, A. O., V. Y. Zitserman, G. A. Kobzev, et A. V. Kosinov. 2019. “Curation of digital scientific data”. *Scientific and Technical Information Processing* 46 (3), 192-203.

- Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymiski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, et Russell D. Gray. 2018. “Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics.” *Scientific Data* 5, no. 1 : 1-10.
- ISO-12620 (2019). *Management of terminology resources – Data category specifications*. Standard, International Organization for Standardization, Geneva, CH.
- ISO-16642 (2017). *Computer applications in terminology – Terminological markup framework*. Standard, International Organization for Standardization, Geneva, CH.
- ISO-30042 (2019). *Management of terminology resources – TermBase eXchange (TBX)*. Standard, International Organization for Standardization, Geneva, CH.
- Kemps-Snijders, Marc, Menzo Windhouwer, Peter Wittenburg, et Sue Ellen Wright. 2008. “ISOCat: Corraling data categories in the wild.” In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.
- McLure, Merinda, Allison V. Level, Catherine L. Cranston, Beth Oehlerts, et Mike Culbertson. 2014. “Data curation: a study of researcher practices and needs.” *portal: Libraries and the Academy* 14, no. 2 : 139-164.
- Palmer, Carole, Nicholas M. Weber, Allen H. Renear, et Trevor Muñoz. 2013. “Foundations of data curation: The pedagogy and practice of “purposeful work” with research data”. *Archives Journal* 3.
- Vezzani, Federica, et Giorgio Maria Di Nunzio. 2019. “Computational terminology in eHealth”. In *Italian Research Conference on Digital Libraries*, pp. 72-85. Springer.
- Vezzani, Federica. 2020. *Vers une “Terminologie FAIR”: conception et implémentation de la ressource TriMED*, Thèse de doctorat, Université de Padoue.
- Vezzani, Federica et Giorgio Maria Di Nunzio. 2020a. “Methodology for the standardization of terminological resources: design of TriMED database to support multi-register medical communication”. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 26 (2), 266-298.
- Vezzani, Federica et Giorgio Maria Di Nunzio. 2020b. “On the Formal Standardization of Terminology Resources: The Case Study of TriMED.” In *Proceedings of the 12th Language Resources and Evaluation*

- Conference, Marseille, France*, pp. 4903-4910. European Language Resources Association.
- Vezzani, Federica. 2021. “La ressource FAIRterm : entre pratique pédagogique et professionnalisation en traduction spécialisée”, *Synergies Italie*, n. 17., p. 51-64.
- Warburton, Kara, et Sue Ellen Wright. 2019. “A data category repository for language resources.” *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, 69.
- Wilkinson, Mark D., Michel Dumontier, Ij J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg *et al.* 2016. “The FAIR Guiding Principles for scientific data management and stewardship.” *Scientific data*, 3(1), 1-9.
- Windhouwer, Menzo, et Ineke Schuurman. 2014. “Linguistic resources and cats : how to use ISOcat, RELcat and SCHEMAcat.” In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3806-3810. European Language Resources Association (ELRA).

Abstract

This study aims at the description of a workflow for the optimal management of terminological data and metadata collected within terminological resources. We focus on the “FAIR Terminology” paradigm aiming to provide guidelines for the provision of findable, accessible, interoperable, and reusable terminology research data. In particular, we focus on the description of a new Repository of Data Categories - designed as information classes - specially implemented in support of the multilingual and multipurpose terminological resource TriMED for the medical domain. The repository is structured according to the guidelines provided in ISO 12620: 2019 and represents one of the steps towards the “FAIRification” of the terminological data contained in TriMED.

