AIUCD 2021

DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale

10° congresso annuale PISA 19-22 gennaio

DIGITAL PUBLIC HUMANITIES

OPEN CULTURE

RETI SOCIALI

TECH ECONOMY

E-PARTICIPATION

TECNOLOGIE ASSISTIVE

Versione PROVVISORIA del contributo presentato al Convegno Annuale

DISCLAIMER

Questa versione dell'abstract non è da considerarsi definitiva e viene pubblicata esclusivamente per facilitare la partecipazione del pubblico al convegno AIUCD 2021

Il Book of Abstract contenente le versioni definitive e dotato di ISBN sarà disponibile liberamente a partire dal 19 gennaio sul sito del convegno sotto licenza creative commons.

On the Reusability of Terminological Data

Giorgio Maria Di Nunzio[1,2], Federica Vezzani[3]

[1] Dept. of Information Engineering, University of Padua, Italy – giorgiomaria.dinunzio(«»)unipd.it
[2] Dept. of Mathematics, University of Padua, Italy
[3] Dept. of Linguistic and Literary Studies, University of Padua, Italy – federica.vezzani(«»)unipd.it

ABSTRACT

Reusability of data is one of the most important practices in science, and investments in this (underestimated) operation may have positive long-term consequences in research (Pasquetto et al. 2017). In this paper, we discuss the benefits of this approach in terminology management by presenting a methodology for the preservation of multilingual terminological records and the practice of standardization as a fundamental step towards reusability. We present a case study to show the effectiveness of this methodology on an obsolete Website containing a multilingual medical glossary, and we share the source code as well as the standardized dataset.

KEY-WORDS

Terminology reusability, terminology management, medical termbases, open science, standardization

1. INTRODUCTION

The design and implementation of high-quality terminological resources is expensive and time-consuming (Warburton, 2015); nevertheless, this costly process is mandatory if the resource has to be reusable, integrated, and useful for a variety of tasks (Schmitz, 2012). Reusability is a keystone in the management of scientific data and, more generally, fits into a wider context of "data curation", as a good practice for organizing, preserving and enhancing research data (Poole, A., 2013). In this regard, the FAIR principles supported by the European Open Science Cloud (EOSC)[1] stress the needs for reusability of data and metadata by means of well-described and standardized approaches (Wilkinson et al., 2016).

In this context, reusability in terminology implies that terminological databases should comply with a standard format, such as the *TermBase eXchange* (TBX) format, promoted by the ISO standard 30042: 2019, so that they can be interoperable and reusable (Melby et al., 2001; Melby, 2015).

In this paper, we focus on a methodology for preserving and enhancing terminological research data in order to open the possibility for developing cross-lingual and multi-lingual applications. We present a case study on 1) the recovery of medical terminological data stored in obsolete HTML pages of a European project; 2) the transformation of these data into the most recent ISO TBX standard; 3) the design and implementation of a new interactive Website to show the potential of the reusability of these data.

2. TERMINOLOGY REUSABILITY

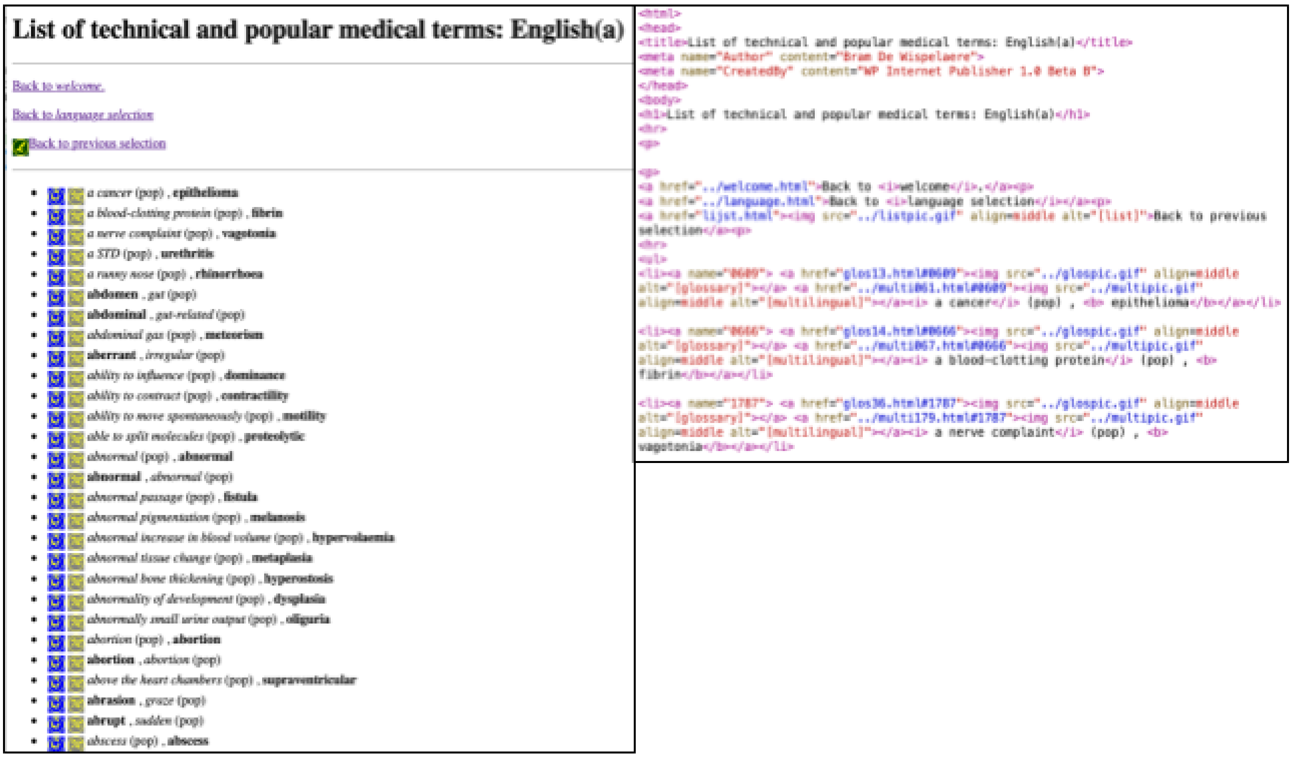
In this section, we describe the methodology for transforming a semi-structured data stored in an obsolete Website into a reusable and standardized multilingual terminological termbase. Our case study is the "Multilingual Lemma Collection" website, which was part of the Multilingual Glossary of technical and popular medical terms.[2] The website, updated the last time on the 3rd of June 2000, contains about 1,800 medical concepts translated in eight[3] of the twenty-four official European languages: English, Dutch, French, German, Italian, Spanish, Portuguese and Danish, both in their technical and popular variants.

2.1 SCRAPING DATA

In Figure 1, we show a portion of the first page of the English monolingual glossary and the corresponding HTML code. The terminological records of the original Web site are stored as items of an unnumbered list, where each item contains the technical term and the popular term separated by a comma, both identified by the value in the "name" tag (for

[1] https://www.eosc-portal.eu
[2] https://users.ugent.be/~rvdstich/eugloss/welcome.html
[3] The project also declares the presence of the Greek language as a working language, but the terms are not available.

example, in Figure 1, 0609 is the identifier for the term *epithelioma*). We use a pipelined R “tidyverse” approach[4] to analyze and extract the data from each webpage, to store and reorganize them into standard XML. The output of this process (source code available online[5]) produces a set of 14,471 terminological entries.



List of technical and popular medical terms: English(a)

Back to welcome.

Back to language selection

Back to previous selection

- *a cancer* (pop) , **epithelioma**
- *a blood-clotting protein* (pop) , **fibrin**
- *a nerve complaint* (pop) , **vagotonia**
- *a STD* (pop) , **urethritis**
- *a runny nose* (pop) , **rhinorrhoea**
- **abdomen** , *gut* (pop)
- **abdominal** , *gut-related* (pop)
- *abdominal gas* (pop) , **meteorism**
- **aberrant** , *irregular* (pop)
- *ability to influence* (pop) , **dominance**
- *ability to contract* (pop) , **contractility**
- *ability to move spontaneously* (pop) , **motility**
- *able to split molecules* (pop) , **proteolytic**
- *abnormal* (pop) , **abnormal**
- **abnormal** , *abnormal* (pop)
- *abnormal passage* (pop) , **fistula**
- *abnormal pigmentation* (pop) , **melanosis**
- *abnormal increase in blood volume* (pop) , **hypervolaemia**
- *abnormal tissue change* (pop) , **metaplasia**
- *abnormal bone thickening* (pop) , **hyperostosis**
- *abnormality of development* (pop) , **dysplasia**
- *abnormally small urine output* (pop) , **oliguria**
- *abortion* (pop) , **abortion**
- **abortion** , *abortion* (pop)
- *above the heart chambers* (pop) , **supraventricular**
- **abrasion** , *graze* (pop)
- **abrupt** , *sudden* (pop)
- *abscess* (pop) , **abscess**

```
<html>
<head>
<title>List of technical and popular medical terms: English(a)</title>
<meta name="Author" content="Bram De Wispelaere">
<meta name="CreatedBy" content="WP Internet Publisher 1.0 Beta B">
</head>
<body>
<h1>List of technical and popular medical terms: English(a)</h1>
<hr>
<p>

<p>
<a href="../welcome.html">Back to <i>welcome</i>.</a><p>
<a href="../language.html">Back to <i>language selection</i></a><p>
<a href="lijst.html"><img src="../listpic.gif" align=middle alt="[list]">Back to previous
selection</a><p>
<hr>
<ul>
<li><a name="0609"> <a href="glos13.html#0609"><img src="../glospic.gif" align=middle
alt="[glossary]"></a> <a href="../multi061.html#0609"><img src="../multipic.gif"
align=middle alt="[multilingual]"></a><i> a cancer</i> (pop) , <b> epithelioma</b></a></li>

<li><a name="0666"> <a href="glos14.html#0666"><img src="../glospic.gif" align=middle
alt="[glossary]"></a> <a href="../multi067.html#0666"><img src="../multipic.gif"
align=middle alt="[multilingual]"></a><i> a blood-clotting protein</i> (pop) , <b>
fibrin</b></a></li>

<li><a name="1787"> <a href="glos36.html#1787"><img src="../glospic.gif" align=middle
alt="[glossary]"></a> <a href="../multi179.html#1787"><img src="../multipic.gif"
align=middle alt="[multilingual]"></a><i> a nerve complaint</i> (pop) , <b>
vagotonia</b></a></li>
```

2.2 STANDARDIZATION AND VISUALIZATION

In order to provide a structured and standardized terminological version of the data recovered from the original HTML, we chose the structure of the terminological record of the TriMED multilingual medical termbase (Vezzani and Di Nunzio, 2020) which refers to the latest ISO standard 30042 (TBX) format (ISO-30042, 2019) and therefore can be reused in any terminology management systems.

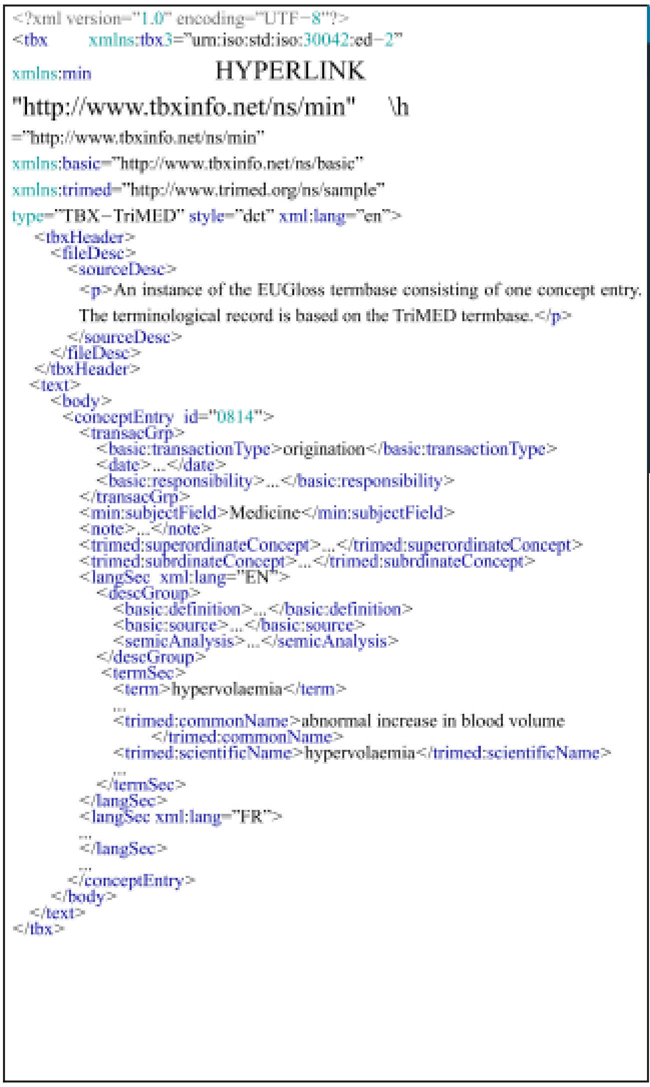
In Figure 2, we show an excerpt of the terminological record generated for the term *hypervolemia* recovered from the Multilingual Glossary. Most pieces of information are missing since the original glossary was a simple one; nevertheless, the missing items can now be filled-in semi-automatically by means of medical ontologies, such as the Unified Medical Language System (UMLS),[6] or manually by translators during their work.

The standardization of the terminological data allows not only the interoperability among different systems, but also an easier reuse in other applications that need these data. As an example, we designed and implemented an interactive Web page[7] that uses the standardized data, based on the following requirements:

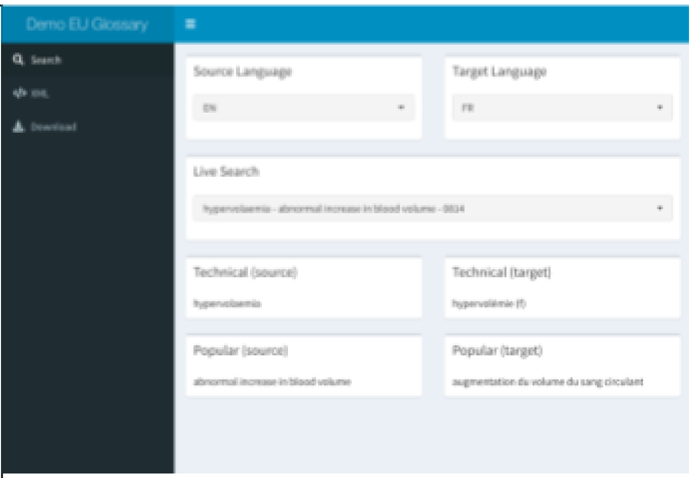
- Group all the main features of the original Web page into a single interactive page,
- Show the standardized XML version of the data,
- Allow the download of the terminological record.

Instead of a static browsing, in this Web application the user can search a term in any of the (source) languages using a keyword completion live search text area. If requested, the user can also choose one of the other (target) languages and see the terminological record as a bilingual glossary with both the technical and popular variants.

[4] https://www.tidyverse.org
[5] anonymized
[6] https://www.nlm.nih.gov/research/umls/index.html
[7] anonymized



```
<?xml version="1.0" encoding="UTF-8"?>
<tbx        xmlns:tbx3="urn:iso:std:iso:30042:ed-2"
xmlns:min                    HYPERLINK
"http://www.tbxinfo.net/ns/min"     \h
="http://www.tbxinfo.net/ns/min"
xmlns:basic="http://www.tbxinfo.net/ns/basic"
xmlns:trimed="http://www.trimed.org/ns/sample"
type="TBX-TriMED" style="dct" xml:lang="en">
   <tbxHeader>
      <fileDesc>
         <sourceDesc>
           <p>An instance of the EUGloss termbase consisting of one concept entry.
           The terminological record is based on the TriMED termbase.</p>
         </sourceDesc>
      </fileDesc>
   </tbxHeader>
  <text>
      <body>
        <conceptEntry  id="0814">
           <transacGrp>
              <basic:transactionType>origination</basic:transactionType>
              <date>...</date>
              <basic:responsibility>...</basic:responsibility>
           </transacGrp>
           <min:subjectField>Medicine</min:subjectField>
           <note>...</note>
           <trimed:superordinateConcept>...</trimed:superordinateConcept>
           <trimed:subrdinateConcept>...</trimed:subrdinateConcept>
           <langSec  xml:lang="EN">
              <descGroup>
                 <basic:definition>...</basic:definition>
                 <basic:source>...</basic:source>
                 <semicAnalysis>...</semicAnalysis>
              </descGroup>
               <termSec>
                 <term>hypervolaemia</term>
                 ...
                 <trimed:commonName>abnormal increase in blood volume
                       </trimed:commonName>
                 <trimed:scientificName>hypervolaemia</trimed:scientificName>
                 ...
              </termSec>
           </langSec>
           <langSec xml:lang="FR">
           ...
           </langSec>
           ...
         </conceptEntry>
      </body>
  </text>
</tbx>
```



Source Language: EN

Target Language: FR

Live Search: hypervolaemia - abnormal increase in blood volume - 0814

Technical (source): hypervolaemia

Technical (target): hypervolémie (f)

Popular (source): abnormal increase in blood volume

Popular (target): augmentation du volume du sang circulant

3. CONCLUSIONS

The design and implementation of multilingual terminological resources open the possibility for developing cross-lingual and multi-lingual applications. In this paper, we addressed the problem of "terminology reusability" as a specific instance of the problem of the implementation of multilingual terminological resources.

We presented a case study and a pipeline for data retrieval, manipulation, and disposal as well as the process of standardization of these records according to the ISO standard in force (TBX) for terminology management based on the TriMED multilingual medical termbase. The standardized machine-readable format allows the reusability of terminological data for multilingual applications as well as the re-design of an interactive Web page that significantly improves user experience for lay-people while looking for medical information.

BIBLIOGRAPHICAL REFERENCES

[1] ISO-30042. Management of terminology resources – TermBase eXchange (TBX). Standard, International Organization for Standardization, Geneva, CH, (2019).

[2] Melby, A., Schmitz, K.-D., and Wright, S. E. Terminology interchange. *Handbook of Terminology Management: Volume 2: Application-Oriented Terminology Management* (2001).

[3] Melby, A. TBX: A terminology exchange format for the translation and localization industry. *Handbook of Terminology*, (2015): 393–424

[4] Pasquetto, I.V., Randles, B.M. and Borgman, C.L. On the Reuse of Scientific Data. *Data Science Journal*, 16, (2017) DOI: http://doi.org/10.5334/dsj-2017-008

[5] Poole, A. Now is the Future Now? The Urgency of Digital Curation in the Digital Humanities. *Digital Humanities Quarterly*, 7 (2) (2013). http://www.digitalhumanities.org/dhq/vol/7/2/000163/000163.html

[6] Schmitz, K.-D. Using international standards for terminology exchange. *Terminologija*, 19 (2012): 33–38.

[7] Vezzani, F. and Di Nunzio, G. M. On the Formal Standardization of Terminology Resources: The Case Study of TriMED. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May 11-16 (2020).

[8] Wachsmuth, H. Text Analysis Pipelines - Towards Ad-hoc Large-Scale Text Mining. *Lecture Notes in Computer Science* 9383, Springer, ISBN 978-3-319-25740-2, (2015): pp. 1-238. https://doi.org/10.1007/978-3-319-25741-9

[9] Warburton, K. Managing terminology in commercial environments. *Handbook of terminology*, 1 (2015): 359– 391.

[10] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3 (1) (2016). https://doi.org/10.1038/sdata.2016.18