The Glossaryfication Web Service – an automated glossary creation tool to support One Health communication

Estibaliz Lopez de Abechuco[‡], Nazareno Scaccia[‡], Taras Günther[‡], Matthias Filter[‡]

‡ German Federal Institute for Risk Assessment (BfR), Berlin, Germany

Corresponding author: Estibaliz Lopez de Abechuco (estibaliz.lopez-de-abechuco-garrido@bfr.bund.de)

Abstract

Efficient communication and collaboration across sectors is an important precondition for true One Health Surveillance (OHS) activities. Despite the overall willingness to embrace the One Health paradigm, it is still challenging to accomplish this in day-to-day practice due to the differences in terminology and interpretation of sector-specific terms. In this sense, simple interventions like the inclusion of integrative glossaries in OHS documents (e.g. reports, research papers and guidelines) would help to reduce misunderstandings and could significantly improve the written communication in OHS. Here, we present the Glossaryfication Web Service that generates a document-specific glossary for any text file provided by the user. The web service automatically adds the available definitions with their corresponding references for the words in the document that match with terms in the user-selected glossaries.

The Glossaryfication Web Service was developed to provide added value to the OHEJP Glossary that was developed within the OHEJP project ORION. The OHEJP Glossary improves the communication and collaboration among OH sectors by providing an online resource that lists relevant OH terms and sector-specific definitions. The Glossaryfication Web Service supports the practical use of the curated OHEJP Glossary and can also source information from other glossaries relevant for OH professionals (currently supporting the online CDC, WHO and EFSA glossaries).

The Glossaryfication Web Service was created using the open-source software KNIME and the KNIME Text Processing extension (https://www.knime.com/knime-text-processing). The Glossaryfication KNIME workflow is deployed on BfR's KNIME Server infrastructure providing an easy-to-use web interface where the users can upload their documents (any text-type file e.g. PDF, Word, Excel) and select the desired glossary to compare with. The Glossaryfication KNIME workflow reads in the document provided via the web interface and applies natural language processing (e.g. text cleaning, stemming), transforming

(bag-of-words generation) and information retrieval methods to identify the matching terms in the selected glossaries.

The Glossaryfication Web Service generates as an output a table containing all the terms that match with the selected glossaries. It also provides the available definitions, corresponding references and additional meta-information, e.g. the term frequency, i.e., how often each term appears in the given text, and the sectoral classification (only for the OHEJP Glossary terms). Furthermore, the workflow generates a tag cloud where the terms are categorized as: (i) exact match when the term in the text matches exactly with the entry of this term in the glossary; (ii) inexact match when the term appears in the text slightly modified (e.g. plural forms or suffixes) and (iii) non-matching that corresponds to all the other words appearing in the text that do not match with any glossary term. Through the user interface, the users can then choose if they want to download the whole list of terms, select only the exact/inexact matching terms, or just choose those terms and definitions that match with the meaning intended for this term in the user-provided document. The resulting table of terms can be downloaded as an Excel file and added to the user's document as a document-specific glossary.

The Glossaryfication Web Service provides an easy-to-adopt solution to enrich documents and reports with more comprehensive and unambiguous glossaries. Furthermore, it improves the referentiality of terms and definitions from different OH sectors. An additional feature provided by the Glossaryfication Web Service is the possibility of extending its use to other glossaries from other national or international institutions allowing the user to customize this glossary creation service.

Keywords

Glossary creation; text processing; One Health Surveillance; communication; collaboration; KNIME

Presenting author

Estibaliz Lopez de Abechuco

Presented at

One Health EJP Annual Scientific Meeting Satellite Workshop 2021 Software Fair

Funding program

Grant title

OHEJP ORION (**O**ne health su**R**veillance **I**nitiative on harm**O**nization of data collection and interpretatio**N)**

Conflicts of interest