The Glossaryfication Web Service: an automated glossary creation tool to support the One Health community

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

‡ German Federal Institute for Risk Assessment, Berlin, Germany

Corresponding author: Nazareno Scaccia (nazareno.scaccia@gmail.com), Matthias Filter (matthias.filter@bfr.bund.de)

Academic editor: Editorial Secretary

Abstract

Background

In many interdisciplinary research domains, the creation of a shared understanding of relevant terms is considered the foundation for efficient cross-sector communication and interpretation of data and information. This is also true for the domain of One Health (OH) where many One Health Surveillance (OHS) documents rarely contain glossaries with a list of terms for which their specific meaning in the context of the given document is defined (Cornelia et al. 2018, Buschhardt et al. 2021). The absence of glossaries within these documents may lead to misinterpretation of surveillance results due to the wrong interpretation of terminology specifically when term definitions differ across OH sectors. Under the One Health EJP project ORION, the OHEJP Glossary was recently created. The OHEJP Glossary is a tool to improve communication and collaboration amongst OH sectors by providing an easy-to-use online resource that lists relevant OH terms and sector-specific definitions. To improve the accessibility of content from the OHEJP Glossary and support the creation of integrative glossaries in future OHS-related documents, the OHEJP Glossaryfication Web Service was created. This service can support the practical use of the OHEJP Glossary and other relevant online glossaries by OH professionals.

New information

The Glossaryfication Web Service (GWS) is an application that automatically identifies terms in any uploaded text-based document and creates a document-specific list of matching definitions in selected online glossaries. This auto-generated document-specific glossary can easily be adjusted by the user, for example, by selecting the desired

definition in case multiple definitions were found for a specific term. The document-specific glossary could then be downloaded, manually adjusted and finally included into the original document where it supports the correct interpretation of terminology used. Especially in sector-specific reports, such as from animal health or public health authorities, this can be beneficial to ensure the correct interpretation by other OH sectors in the future. The GWS was developed with the open-source desktop software KNIME Analytics Platform and runs as a web service on a KNIME Web Server infrastructure. The core data processing functionality in the GWS is based on KNIME's Text Processing extension. KNIME's JavaScript nodes provided the basis for an interactive user interface where users can easily upload their files and select between different reference glossaries, such as the OHEJP Glossary, the CDC Glossary, the WHO Glossary or the EFSA Glossary. After retrieval of the user input settings, the GWS tags words within the provided document and maps these tagged words with matching entries in the selected glossaries. As the main output, the user receives a downloadable list of matching terms with their corresponding definitions, sectorial assignments and references, which can then be added by the user to the original document. The GWS is freely accessible via this link as well as the underlying KNIME workflow.

Keywords

One Health, OHEJP Glossary, text processing, KNIME, glossary creation

Introduction

The One Health (OH) approach aims to improve the collaboration across multiple disciplines at the national and international level in order to prevent and control emerging zoonotic diseases, as well as antibiotic resistance (Bordier et al. 2018, Cornelia et al. 2018). The practical implementation of the OH paradigm is therefore closely linked to the challenge of integrating information from different domains and sectors, such as food safety, public health, animal health and environmental sciences. Great collaborative efforts are currently made by the OH community and this integrated form of surveillance is defined as One Health Surveillance (OHS) (Bordier et al. 2019, Bordier et al. 2018, Cornelia et al. 2018). Despite the efforts undertaken to promote cross-sectoral collaboration, differences across sectors in how terminology is used and interpreted remain a significant barrier (Buschhardt et al. 2021). The international and cross-sectorial work, previously conducted within the ORION project, has shown that different countries, as well as different sectors, have established their own terminology and definitions that might lead to misunderstandings when exchanging information across sectors. The risk of ambiguous communication increases in this OH context where experts from different sectors work together (Buschhardt et al. 2021). The use of glossaries can facilitate the integration of information arising from different OH disciplines and lower the barrier related to the communication. For instance, OHS documents, such as reports or guidelines, not always contain glossaries with definitions for all OHS-related terms. As a result, misinterpretation of data and results from sector-specific reports can occur. The

incorporation of comprehensive glossaries within all OH-related documents can help the readers to not interpret the meaning of terms differently from the authors. The first step towards a clearer definition of OH-related terms from different OH sectors was the recently-established OHEJP Glossary (Buschhardt et al. 2021).

Here, we presented a new web-based solution named Glossaryfication Web Service (GWS) that enables users to automatically create document-specific glossaries by automatically searching through the OHEJP Glossary and other international glossary resources, such as CDC (Centers for Disease Control and Prevention), EFSA (European Food Safety Authority) and WHO (World Health Organisation). The GWS was designed as a user-friendly tool that adds additional value to the OHEJP Glossary and provides an easy approach to enrich future OH-related documents with more comprehensive and unambiguous glossaries. It was implemented using the open-source software KNIME (Berthold et al. 2009) and its Text Processing extension (Thiel 2009). Further, it was designed such that it can be easily extended to make use of other online glossaries from national or international institutions. For example, the GWS was customised and deployed as an independent web service for German glossaries.

Project description

Title: The Glossaryfication Web Service: an automated glossary creation tool to support the One Health community

Design description: The GWS is an easy-to-use web-based tool to generate a document-specific glossary during document writing. The GWS automatically identifies nouns in a user-provided text document and yields matching definitions from the OHEJP Glossary and other supported online glossaries. Users can upload documents in different formats (e.g. PDF, Word, Excel) and choose one or more glossary(ies) to search through. Next, the GWS will automatically search within the user-provided text document for terms that are contained in the selected glossary(ies). The user will receive a downloadable list of matching glossary terms with their corresponding definitions, which can be added to the user's document. The GWS produces a tag cloud and a table of terms with all found definitions. This table reports on exact and partly matching terms, as well as all terms for which no matching entry could be found in any of the selected online glossaries. The exact matches refer to those terms equally written in the user's provided text document and the reference glossary(ies). Inexact matches are those terms that do not exactly match the reference glossary terms, but match with small changes (e.g. plural form). The GWS provides these results in an interactive dashboard, where the end-users can select those terms and definitions that best match the intended meaning within the user-provided text document. The provided list of matching terms contains the definitions, along with their sectoral classification, references and information on the term's abundance in the provided text document. This table can easily be downloaded as an Excel file and then edited further to be finally added to the user's document as a glossary. The GWS is accessible via regular web browsers and is operated on BfR´s KNIME Server WebPortal. On the start page of the service, a brief description is provided (Fig. 1). The

end-user can upload the own report, select the appropriate reference glossary(ies) and, in the end, download the generated list of matching terms with the corresponding definitions and reference information. In Fig. 1, a step-by-step description of the GWS is given.

Funding: This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 773830.

Web location (URIs)

Homepage: https://foodrisklabs.bfr.bund.de/ohejp-glossary/

Technical specification

Platform: Web browsers

Operational system: Windows; Linux; Mac

Usage licence

Usage licence: Other

Implementation

Implements specification

The GWS infrastructure was created with the open-source software KNIME (Konstanz Information Miner) Analytics Platform, which is a user-friendly graphical workbench for data science (Berthold et al. 2009). With the KNIME platform, the construction of a data pipeline (workflow) is possible by the connection of modules (nodes) in a way that the output of a node is the input of the subsequent one. Each KNIME node represents a specific data processing module where the collection of KNIME nodes are able to perform all sorts of data-related tasks (e.g. reading/writing files, data transforming, data visualisation etc.) (Berthold et al. 2009). Each KNIME node encloses all its required functionalities so that a KNIME workflow, composed of different nodes, can store the whole data analysis process including all intermediate results. This KNIME modular principle ensures an easy extension, adaptation and implementation of new functionalities to the already existing KNIME workflows (Berthold et al. 2009). The GWS KNIME workflow consists of component nodes for the different web-based user interfaces (start page with service description and user input; interactive dashboard with results; page to download selected entries as Excel files) and five data processing steps (data integration, text processing, data preprocessing, combine matches and HTML table view

preprocessing) (Fig. 2 A). In the first step (data integration), reference glossaries, such as the OHEJP Glossary and other resources obtained from international institutions, were used. The reference glossaries of the international agencies used were downloaded from the glossary section of their respective websites (CDC, EFSA and WHO), joined in a single Excel file and, subsequently, uploaded within the workflow. The OHEJP Glossary is the only glossary where the GWS can automatically retrieve updates as it is equipped with an API. In the second step (text processing), the workflow reads in textual data provided by the end-user and applies natural language processing, text mining and information retrieval methods, based on the KNIME's Text Processing extension (Thiel 2009) to identify matching terms in the reference glossaries. The KNIME's Text Processing extension is generally divided into different categories such as IO, enrichment, preprocessing, transformation and frequencies; each of them harbouring several nodes with a specific function (Fig. 2 B). The text processing starts with the Tika Parser node (Fig. 2 B), which reads in textual data into KNIME and extracts textual contents and metadata from different file formats. After the text is uploaded, the textual data provided are enriched by named-entity recognition and tagging. Particularly, the Dictionary Tagger node (Fig. 2 B), recognises named entities specified in a reference dictionary and assigns a specified tag type and value. This node was configured either for an exact match or case-sensitive named entity recognition as tagger options. In addition, the OpenNLP English WordTokenize was selected as a word tokeniser. Afterwards, the preprocessing metanode removes stop words, numbers and punctuation markers, so that the number of irrelevant words is reduced and the performance of the workflow is increased (Fig. 2 B). Within the preprocessing metanode, the Snowball Stemmer node for stemming algorithms is also included. The stemming reduces the word to its word stem and is used for the identification of inexact matches (e.g. environmental or environment). Next, under the transformation and frequencies step, the creation of a bag of words and computation of the term frequency is performed (Fig. 2 B). Subsequently, the terms obtained in the bag of words are further compared with the reference glossary(ies) in order to retrieve exact matches. The string matching performed by the String Matcher node of stemmed terms within the metanode identifies the inexact matches. Subsequently to the text processing step, exact-, inexact- and non-matching terms are further preprocessed. Exact matches and inexact matches are separated according to the tag value generated in the text process steps and compared with the reference glossary(ies). For inexact matches, the Levenshtein distance (function to calculate a number of modifications needed to change one word into another) is computed. In step four of the Glossaryfication workflow, all the matches are combined in a single table. In the last step, the combined table is joined with the reference glossary(ies) content in order to retrieve reference links and sector classification of the matches found. The HTML table view of the dashboard is also generated in this step (Fig. 2 A). The GWS KNIME workflow can be executed in two manners: locally on a desktop PC after KNIME software installation; or directly online using the KNIME WebPortal via a web browser. The latter is possible through the deployment of the GWS KNIME workflow on the BfR's KNIME Server infrastructure. This allows the execution of the GWS workflow in an interactive way via a web interface from a web browser without the need to install KNIME by the user.

Additional information

Validation of the Glossaryfication tool and usage

The GWS was also validated using a customised text containing known terms (Suppl. material 1). The validation text used was the following:

"The following terms are found in the WHO reference glossary: **Best available evidence**, *Bbcest available evidence*, **Capital investment**, *Capital investments*, **Commissioning services**, **Community participation**, **Comprehensive (maxi) HIA**, **Concurrent HIA**, **Decision making**, **Determinants of health**, **Disadvantaged / vulnerable / marginalized groups**, **Economic impact assessment**, **Employment Zone**, **Environmental impact assessment**. Actually, some of the above terms are wrongly written."

The terms written in bold were taken from the WHO glossary and were used as "exact matches" to be retrieved by the GWS. The terms written in italic instead, represent "inexact matches" written differently on purpose for the validation exercise. All the other words from the validation text example that contain more than three letters and are not exact or inexact matches, are defined as "non-matching terms". The validation text was uploaded to the GWS via web browser and then the WHO-Glossary was selected as the reference glossary. The service found 13 exact matches out of 12. The additional exact match found is "impact assessment" that is present within the WHO reference glossary. The service finds the term "impact assessment" because it is partially identical to the WHO glossary entry "Economic impact assessment". Indeed, "impact assessment" is also recognised from the service as an inexact match. If, within a text, there is one term which is also part of a compound word that is included in the reference glossary, the compound word will appear as result of the GWS although it is not present within the text. The compound word is recognised from the service as an inexact match. The plural form of the term "Capital investment" is also retrieved as an inexact match with a Levenshtein distance of 1 due to insertion of the letter "s". The compound "Bbcest available evidence" as the inexact match is not found. This can be explained due to the insertion of more than one wrong letter in the term. The term "Bbcest available evidence" is rather divided into three different terms which appear separately as non-matching terms. If terms within the text are misspelled, those will appear as non-matching terms. In total, 27 out of 11 are obtained as non-matching terms. The high number of non-matching terms found is due to inexact matches and to the stemming process. Each word variant of an inexact match compound is reported in the list of non-matching terms. Additionally, these words, together with the other terms, are repeated twice in the list of the non-matching terms as stemmed and non-stemmed words. Examples of customised glossary tables obtained by the GWS have already been published (Filter et al. 2021, Buschhardt et al. 2021).

Acknowledgements

This work was supported by funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 773830: One Health European Joint Programme.

References

- Berthold M, Cebron N, Dill F, Gabriel T, Kötter T, Meinl T, Ohl P, Thiel K, Wiswedel B (2009) KNIME - the Konstanz information miner. ACM SIGKDD Explorations Newsletter 11 (1): 26-31. https://doi.org/10.1145/1656274.1656280
- Bordier M, Uea-Anuwong T, Binot A, Hendrikx P, Goutard FL (2018) Characteristics of One Health surveillance systems: A systematic literature review. Preventive Veterinary Medicine 181: 104560. https://doi.org/10.1016/j.prevetmed.2018.10.005
- Bordier M, Delavenne C, Nguyen DTT, Goutard FL, Hendrikx P (2019) One Health Surveillance: A Matrix to Evaluate Multisectoral Collaboration. Frontiers in Veterinary Science 6 https://doi.org/10.3389/fvets.2019.00109
- Buschhardt T, Günther T, Skjerdal T, Torpdahl M, Gethmann J, Filippitzi M, Maassen C, Jore S, Ellis-Iversen J, Filter M (2021) A one health glossary to support communication and information exchange between the human health, animal health and food safety sectors. One Health 13 https://doi.org/10.1016/j.onehlt.2021.100263
- Cornelia A, Amaia A, Alessandro C, Orlando C, Massimo C, Karl E, Graham F, Celine G, Joseph J, Dominique M, Paul R, Jan S, Jonathan S, Judit T, Svetla T, Emma W, Johanna Y (2018) Towards One Health preparedness. ECDC.
- Filter M, Buschhardt T, Dórea F, Lopez de Abechuco E, Günther T, Sundermann E, Gethmann J, Dups-Bergmann J, Lagesen K, Ellis-Iversen J (2021) One Health Surveillance Codex: promoting the adoption of One Health solutions within and across European countries. One Health 12 https://doi.org/10.1016/j.onehlt.2021.100233
- Thiel K (2009) The KNIME Text Processing Plugin. CiteSeerX. URL: https://www.knime.com/sites/default/files/KNIME-TextProcessing-HowTo.pdf

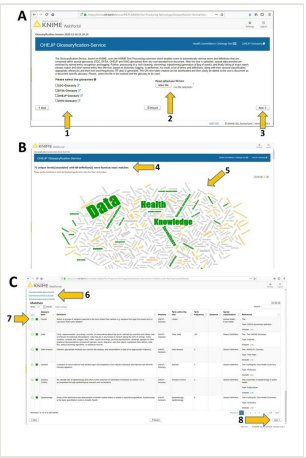

Figure 1.

A, a screenshot of the GWS start page. **B** and **C**, output sections of the service. Specifically, **B** , number of exact terms and associated definitions found and, a tag cloud outcome; **C**, downloadable table of identified terms and definitions with term frequency, sector classification and reference. Through the GWS, end-users can upload their own file (2, A) and select which of the supported glossary(ies) should be searched through (1, A). Clicking “Next” (3, A), the GWS displays the results in an interactive table within the dashboard providing also the number of occurrences for each term in the user's document next to the identified definitions (4, B) and the tag cloud (5, B). The different colours for the terms in the tag cloud refer to different types of matches: exact matches in green, inexact matches in yellow and non-matching terms in grey. Afterwards, scrolling down the dashboard view (image C), the user has the possibility to download directly different tables (6, C) in Excel format. These options are: i) download the table with all the matches found, ii) download the exact-matches table or iii) download the inexact-matches table (6, C). Furthermore, if only a few terms with appropriate definitions are needed, the end-user can download those specific terms checking the corresponding checkboxes (7, C) and then clicking "Next" at the bottom of the dashboard page (8, C).

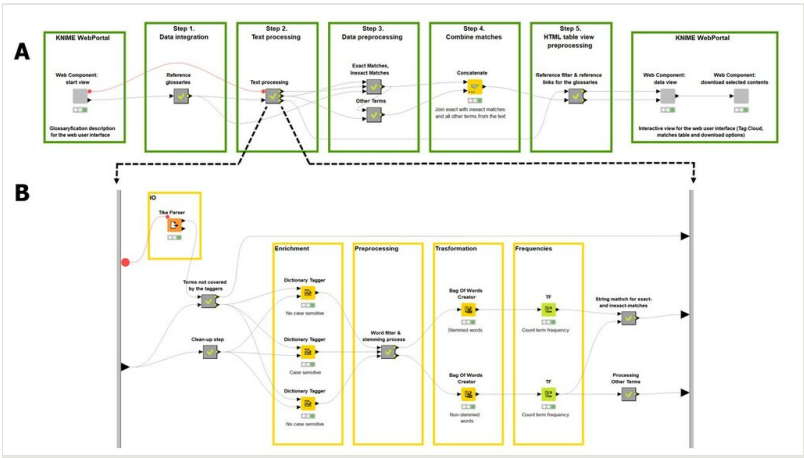

Figure 2.

A, a screenshot of the GWS KNIME workflow. There are 5 steps within the data processing workflow, which are described in the text. The first and the last green boxes of the workflow, designed with KNIME WebPortal extension, contain so-called "Components" that provide a workflow-specific web user interface that can also be triggered by the KNIME Server. In this way, the GWS KNIME workflow becomes available as a fully functional web service in the KNIME WebPortal. **B**, a screenshot of the KNIME's Text Processing extension nodes wrapped up into a metanode.

Supplementary material

Suppl. material 1: The Glossaryfication Web Service: an automated glossary creation tool to support the One Health community

Authors: Nazareno Scaccia, Taras Günther, Estibaliz Lopez de Abechuco, Matthias Filter
Data type: Glossaryfication Web Service workflow
Brief description: Herein, the Glossaryfication Web Service workflow with results is provided. The workflow is already executed using a brief customised text, as described in the additional information section.
Download file (716.74 kb)