Frame-based terminography: a multi-modal knowledge base for karstology

Špela Vintar[1], Vid Podpečan[2], Vid Ribič[3]

[1] University of Ljubljana, Aškerčeva 2, SI – 1000 Ljubljana
[2] Jožef Stefan Institute, Jamova 39, SI – 1000 Ljubljana
[3] Kofein dizajn, Beethovnova 9, SI – 1000 Ljubljana
E-mail: spela.vintar@ff.uni-lj.si , vid.podpecan@ijs.si, vid@kofein.si

Abstract

We present an innovative approach to the representation of domain-specific knowledge which combines traditional concept-oriented terminography with knowledge frames and augments linguistic data with images, videos, interactive graphs and maps. The interface is simple and intuitive, prompting the user to enter a query term in any of the three languages (English, Croatian and Slovene). If the term is found it is described through textual definitions from various sources, its frame derived from annotated data, a graph depicting the neighbourhood of the concept and – if feasible – a map of geolocations for the queried term. The frame represents aggregated and structured knowledge as it describes the concept through a set of semantic relations. Graphs enable the user to browse through related concepts and explore the domain in a visually represented network. The underlying knowledge base of karstology was created within the TermFrame project and is based on an implementation and extension of the frame-based approach to terminology.

Keywords: frame-based terminography; karstology; knowledge base; visualisation

1. Introduction

The notion of frames as templates of knowledge structures (Faber, 2009; Faber et al., 2011) has found great resonance in the field of terminology as it efficiently combines the textual, contextual and cognitive layers of knowledge into a comprehensive theoretical and practical framework. In the TermFrame project we approach the domain of karstology from an interdisciplinary perspective to create a multilingual and multi-modal interactive knowledge base tailored to different types of users: domain experts, students, and researchers, but also non-experts interested in karst.

Karstology itself is an interdisciplinary field studying karst, a special type of landscape which develops on soluble rocks such as limestone or gypsum. Typical karst landmarks include caves, sinkholes, various rock formations and complex water systems with streams which may sink and continue their flow subterraneously. Apart from being a field of interest for geography, hydrology, speleology and geology, karst systems – especially caves – are popular tourist destinations and important areas of environmental protection, which is why we envisage interested non-experts as potential users of our knowledge base.

The web user interface to the knowledge base is designed in line with the principles of usability as defined by Jakob Nielsen through the following five key features (Nielsen, 1996): *learnability* (how simple the interface is for a first-time visitor), *efficiency* (how quickly the user can complete their task), *memorability* (how well does the user master the interface after a period of non-use), *errors* (the number of errors the user makes during use, their gravity and difficulty of correction), and *satisfaction* (how pleasing the interface design is).

The remainder of this article is structured as follows: After a brief overview of related work in Section 2 we dedicate Section 3 to the various sources of information for our knowledge base. We describe the resources, processing steps and tools used to create each of the layers presented to the user. Section 4 focuses on the mode of presentation itself and the rationale of designing the search interface so that it can be accessible and usable for all of our potential target groups. We conclude with a brief discussion and plans for future work.

2. Related work

Frame-based approaches to terminology (FBT; Faber, 2012) have become mainstream in the past decade. While the EcoLexicon as the first of its kind continues to improve and expand (Faber et al., 2016; León-Araúz et al., 2019), other authors and projects integrate frames or conceptual templates into their knowledge representations (Roche et al., 2019; Bihua et al., 2020; Giacomini, 2018).

Since specialised knowledge is often conceptualised as a network, numerous examples of knowledge visualisations in the form of graphs can be listed, such as multilingual databases of colexification patterns CLICS [1] (Mayer et al., 2014), Wikipedia visualisation (WikiGalaxy[2]) or biological domain knowledge exploration software such as Biomine Explorer (Podpečan et al., 2019). The latter implements a rich network visualisation and manipulation interface which sits on top of the Biomine search engine serving the relevant parts of the enormous Biomine network according to the user's query. Cytoscape (Shannon et al., 2003) is one of the most important examples of feature-complete network analysis software. While it was originally developed for biological research, it has since grown into a general, extensible platform for complex network analysis and visualisation. Gephi (Bastian et al., 2009) implements very efficient algorithms for the visualisation of extremely large networks, but does not implement many data integration options and is thus limited to visualisation and basic analysis of general networks. OmicsNet (Zhou and Xia, 2018) implements a visual analytics platform for multi-omics integration and features 3D visualisation in the browser.

[1] http: //clics.lingpy.org

[2] http://wiki.polyfra.me/

3. Resources for the TermFrame knowledge base

3.1 Concepts and textual definitions

The creation of our trilingual knowledge base for karstology was performed in stages (cf. Vintar et al., 2019). First, specialised corpora in English, Croatian and Slovene were compiled, ensuring optimal coverage of the domain. The corpora are comparable and contain relevant contemporary works on karstology, including books, articles, doctoral and master's theses, glossaries and encyclopaedia. The composition of the corpus is described in more detail in Vintar and Stepišnik (2020). The English subcorpus contains just under two million words, while the Slovene and the Croatian subcorpora are smaller and together consist of around one million words.

Some of the corpus texts were available only in printed format, so that a full digitisation procedure was required, including scanning, OCR and manual proofreading; others were obtained directly from publishers, authors and internet sources. For some of the texts copyright issues remain unresolved, and such texts were used only as a source of definitions and their digitised versions have been discarded. The cleared part of the comparable corpus will be released through the Clarin.si repository[3].

In the second stage, definitions of karst concepts were collected from the TermFrame corpora using the ClowdFlows definition extraction tool (Pollak et al., 2012). The final data set consists of 725 annotated definitions for English, 786 for Slovene and 661 for Croatian. All definitions were manually annotated in accordance with our domain model specifying the semantic categories and relations relevant for karstology (cf. Vintar et al., 2019). An example of an annotated definition can be seen in Figure 1.

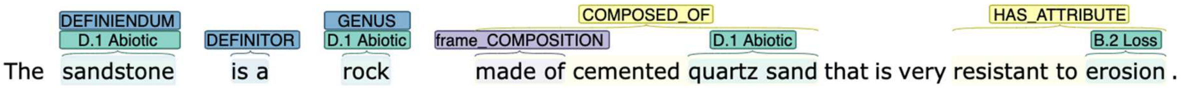

The sandstone is a rock made of cemented quartz sand that is very resistant to erosion .

Figure 1: Annotated definition in WebAnno

The domain model specifies five top-level categories dividing karst terms into Landforms, Processes, Geomes, Entities/Properties and Instruments/Methods. Each category is associated with a set of semantic relations used to define or describe it; these combinations can also be referred to as definition templates and help organise and represent knowledge in a systematic manner. In addition to the categories and relations, each definition is also analysed for definition elements, so that we annotate the DEFINIENDUM, GENUS and SPECIES (the latter is relevant for extensional definitions).

[3] https://www.clarin.si/repository/xmlui/

The definitions in all three languages are contained in a common database where each definiendum – which can be in English, Slovene or Croatian – is assigned a concept ID, thus linking equivalents to a specific and unique meaning. A concept may have several definitions in one language (most notably karst, for which there are as many as 13 English definitions) and several terms designating it, or it may not have an equivalent in all three languages.

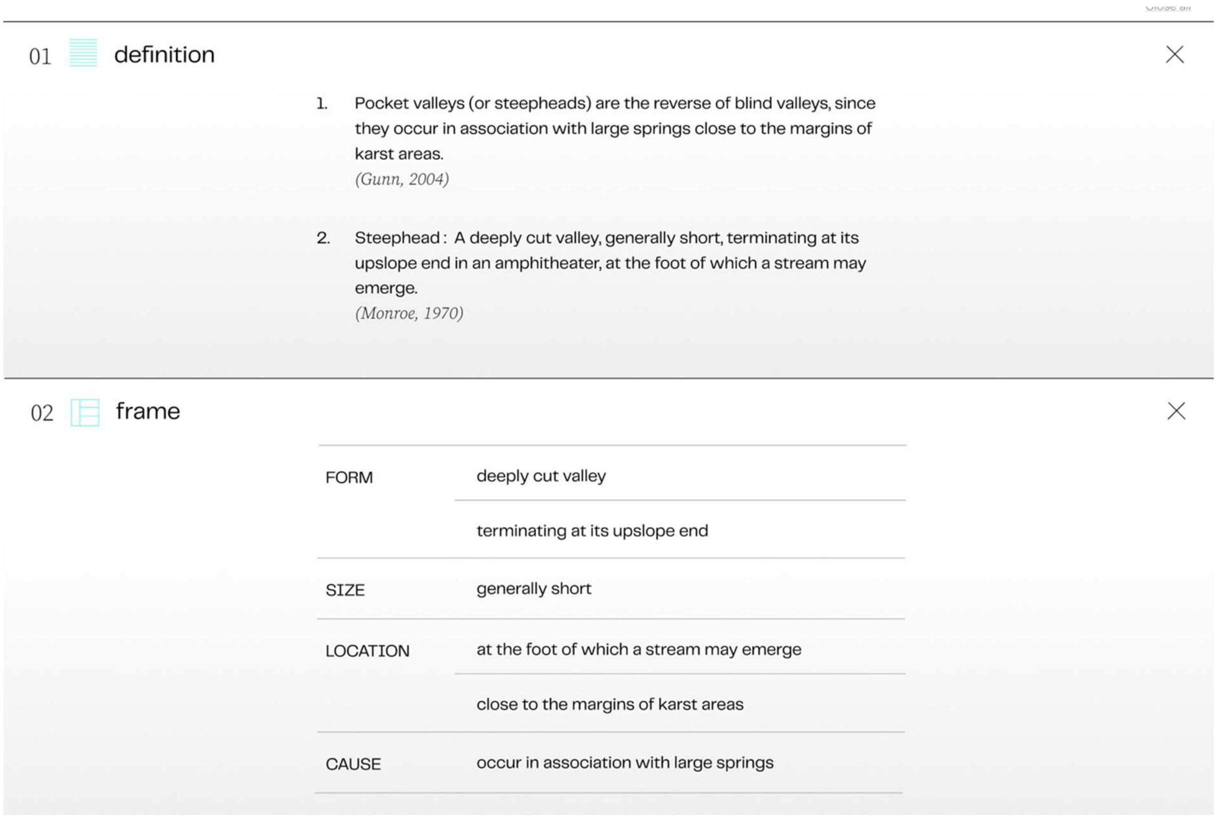
3.2 Representing frames

In order to allow further processing of annotated definitions in all three currently supported languages they have to be exported from the WebAnno annotation software (Eckart de Castilho et al., 2016). We use the common .tsv format which is one of the available outputs of WebAnno. Due to the complexity of the annotated data, any simple text format (including .csv) is ill-suited for this task. The following issues need to be handled by the parser in order to extract the correct and complete data.

- The annotation of a text is composed of annotation blocks which contain annotations of sentences. These blocks are separated by empty lines and comments.
- Single cells may contain additional inner separation characters.
- An annotation can span any number of cells in the same column, either in a contiguous block or possibly separated with other annotations.
- Annotations spanning multiple tokens are characterised by annotation serial numbers (counters) in square brackets following the annotation name. However, serial numbers are not present in annotations spanning single tokens.

We implemented the parser using the popular Pandas framework,[4] which offers several data manipulation and selection features which made our task easier. First of all, the csv parser is configured so that the .tsv export of WebAnno is stored correctly into an internal data structure (Pandas' DataFrame). Then, the complete annotation data is split into sentence annotation blocks using sentence ID as the grouping key. The possibility of intra-cell separation is handled next by duplicating the row for each such value while assigning a new, unique index. This is followed by extracting the actual tokens belonging to each annotation. Pandas' powerful data selection functions are used to simplify this task. Finally, the complete annotation data is stored in an internal format and ready to be converted into a format suitable for the representation of frames in a table or visualisation in a graph.

[4] https://pandas.pydata.org/



01 definition

1. Pocket valleys (or steepheads) are the reverse of blind valleys, since they occur in association with large springs close to the margins of karst areas.
   (Gunn, 2004)

2. Steephead : A deeply cut valley, generally short, terminating at its upslope end in an amphitheater, at the foot of which a stream may emerge.
   (Monroe, 1970)

02 frame

FORM	deeply cut valley
	terminating at its upslope end
SIZE	generally short
LOCATION	at the foot of which a stream may emerge
	close to the margins of karst areas
CAUSE	occur in association with large springs

Figure 2: Definitions and frame

The data presented as the frame of the query term collects all annotated semantic relations from different definitions and displays them in the order of the "ideal" definition. Thus, if Surface landforms are typically defined through their FORM, SIZE, LOCATION and CAUSE, the frame tab will list all strings from the definitions that had been annotated as either of these relations. The main added value of the frame-based approach is that the information about the term is aggregated from different textual sources, and that it is structured in a manner which reflects the cognitive template surrounding the Surface landform concept category.

3.3 Visualisation

There are several possibilities how to define a graph structure using the extracted annotations. Currently, graphs are created according to the following rules applied to each sentence annotation block.

1. For every "definiendum" definition element create:
   a. a node from its tokens,
   b. a node from its category, and
   c. a directed edge named *has_category* from the token node to the category node.

2. For every “genus” node create a node from its tokens.
3. For every “definiendum” token node and every “genus” token node create a directed edge named *is_a* from the first node to the second node.
4. For every “relation” definition element create:
   a. a node from its tokens,
   b. a directed edge from the “definiendum” token node to the “relation” token node, and give it the name of the relation.

The visualisation backend software stack consists of the following components. First, the data loader provides fast loading from serialised data structures containing the graphs with the topology as described above. Second, the graph extraction component performs subgraph extraction according to input parameters. Currently, one or more nodes can be used as the input query. The extractor performs neighbourhood search from the specified nodes using the currently default depth limit of 2 and returns the resulting subgraph. Finally, the exporter serialises the extracted subgraph into a selected format. We use JSON to pass the subgraph data to the frontend, but several other formats are supported and can be used for server-side processing or for download.

The visualisation of the graph corresponding to the user query is implemented using the open source vis.js library[5] which is a dynamic, browser based visualisation library. It enables interactive and efficient visualisation of reasonably large graphs (up to a few thousand nodes). In our case, however, the size of graphs is limited to only few dozens of nodes because of the neighbourhood search depth limit of 2.

When the JSON containing the graph data is received from the backend, a vis.js DataSet structure is created first. It contains information about nodes and edges and any additional node and edge data that is required by the graph visualisation user interface. Then, a visualisation canvas is created and populated with the contents of the DataSet. Several visualisation parameters are set to values which enable clear visualisation of small knowledge graphs.

The graph displayed alongside the query is interactive in the sense that each node which corresponds to a term in our knowledge base can be clicked by the user. This action runs a new query so that the entire results window is refreshed and a new set of definitions, frame, graph, etc. is displayed.

3.4 Images and videos

Since most of our karst concepts pertain to tangible landscape entities, we obtained a collection of images and videos depicting karst phenomena. Images are labelled with concept IDs and integrated into the search interface. Images and aerial photographs of

[5] https://visjs.org/

karst forms and processes were obtained during systematic field surveys and morphographic mapping for documentation and field research of karst conducted by Dr. Uroš Stepišnik and colleagues from 2006 to 2021. Apart from karst documentation and research, the visual materials are also used for didactic purposes in teaching the physical geography of karst at the Department of Geography at the University of Ljubljana (Stepišnik, 2020). Classical photographic equipment and unmanned aerial vehicles were used for photographic documentation. The image and video material is available for the purposes of the TermFrame project under the CC-BY-NC-ND license.

3.5 Maps

For the most central and frequent karst landforms which are described in our corpus through actual geolocations, we created maps displaying these locations. Place names were automatically extracted using the GeoNames.org database as a source of global geographical names and REZI[6], a publicly available registry of geographical names for Slovenia and Croatia. The extracted names were supplemented with GPS coordinates and imported into Google MyMaps to create maps of documented locations of the relevant landform.

4. Designing the interface

The search interface is designed to be as simple and user-friendly as possible, focusing primarily on usability for non-linguists. The user can enter a karst term in any of the three languages and the results will be displayed in tabs. After the image or video, the user can read all the definitions for the concept from different sources, then view the "framed" definition, browse a clickable graph of related terms and, if available, see the locations of the concept on the map.

The web interface is a WordPress installation with some custom modifications tailored to the needs of our project. We have developed a database importer in order to easily import new terms into the website. The importer processes the entries from a csv file and maps them to the corresponding posts in a WordPress database. On top of that we have a cron job which obtains the data for the graph visualisation via API. A cron job is a simple software utility that schedules tasks to run at certain time intervals in the future. This API is specifically developed for this project and returns information about nodes and edges for any karst concept we have in the database. This data is subsequently processed so it can be used with a vis.js library to display the graph in the frontend.

[6] https://egp.gu.gov.si/egp/?lang=en

4.1 Target audience

The first step in designing the user experience is to define the target users of the interface. In our case the interface addresses several target groups of experts and non-experts. Experts from the domains of geography, karstology and speleology will be able to consult the karst knowledge base during their work, compare definitions by different authors and browse for similar concepts. Linguists and terminologists will explore mainly the linguistic aspects of the terms and their definitions, and both groups will benefit from the equivalents in other languages, related concepts and graphs thereof.

The more general target group of non-experts will explore karst phenomena through images, videos, textual descriptions and maps.

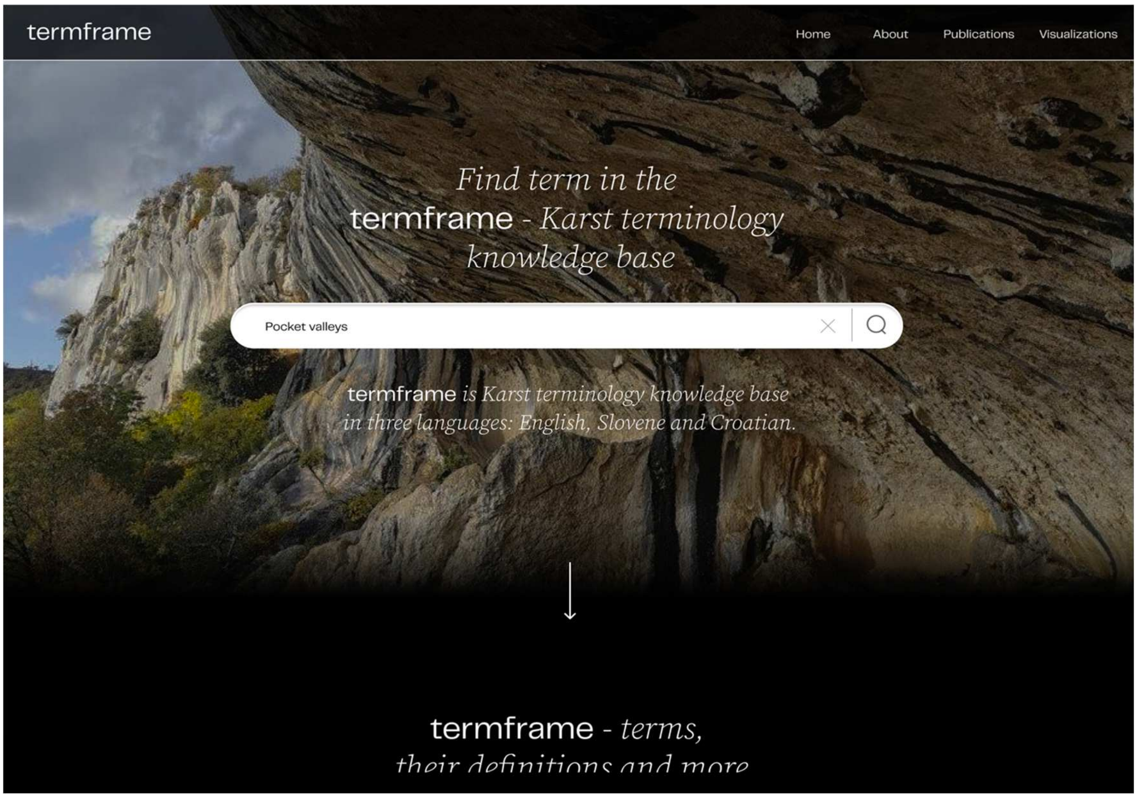
4.2 Browse vs. search

While designing the user interface we first needed to resolve the question of how to represent the knowledge base to facilitate user access and satisfy the five Nielsen criteria of usability mentioned above. The choice was between two user scenarios, browse or search, whereby each has its advantages and disadvantages. Browsing allows the user to search through a list or hierarchy. If the list is unordered, the search time increases linearly with the number of items to choose from. Since our knowledge base contains over 1,700 terms, such browsing would be extremely inefficient.

4.3 User journey

All target groups share the same mode of access to the knowledge base, but upon receiving a response to the query the user may select the most relevant type of content presentation. First, the user enters a query into the search field (for example “pocket valleys”, see Figure 3).

We therefore selected searching as the access scenario. The user enters a query term and immediately receives hits – provided the knowledge base contains the query term. Browsing can be resumed via the graph of related terms where the number of items to choose from is considerably smaller, while still allowing the user to explore without knowing exactly what to search for. The search engine will first display the results in the language of the query term, but the user may switch languages if the same concept is described in the other two languages.

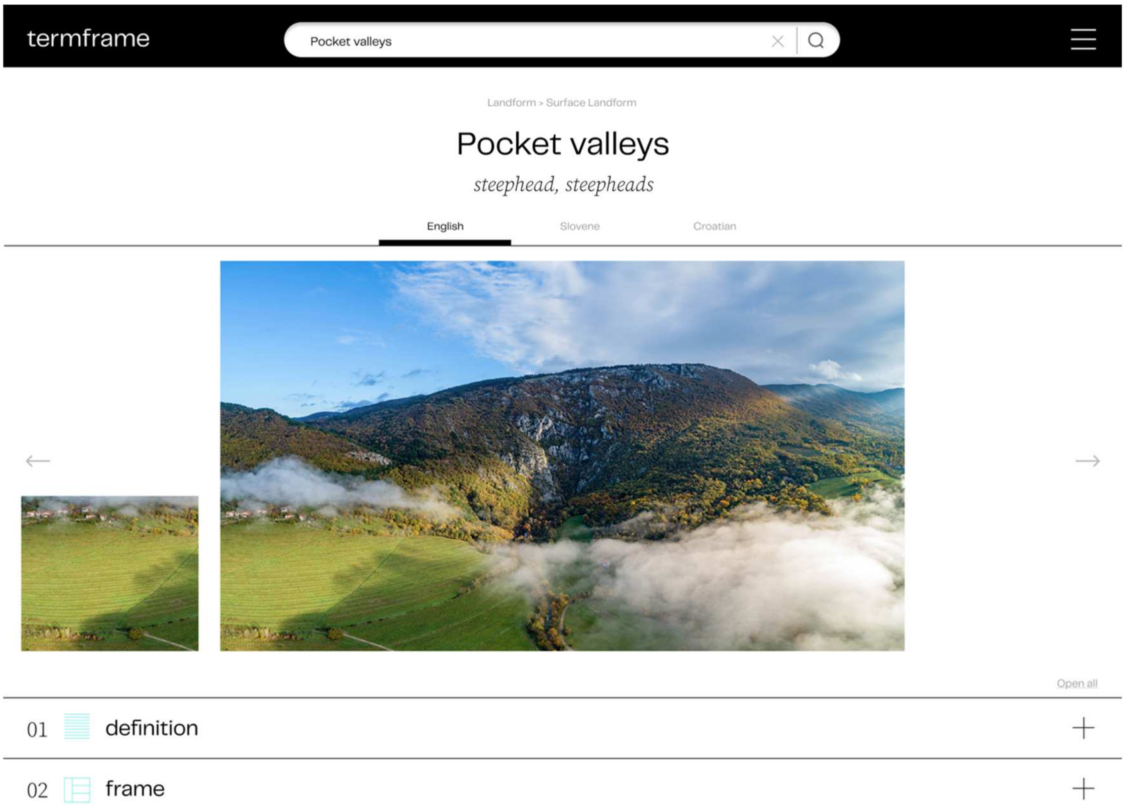


Find term in the
termframe - *Karst terminology knowledge base*

termframe *is Karst terminology knowledge base in three languages: English, Slovene and Croatian.*

termframe - *terms, their definitions and more*

Figure 3: Main search field



Pocket valleys

steephead, steepheads

01 definition

02 frame

Figure 4: Displaying the found term

The query term is displayed in the header of the page, together with the semantic category and subcategory above the term and its synonyms below it (Figure 4). Under the image or video is a list of four expandable tabs for the user to choose from. A domain expert will presumably focus on the definitions and the frame (Figure 2), a linguist might explore the graph (Figure 5), and a non-expert user might open the map (Figure 6) and look for locations of the karst phenomenon.

The web site contains two additional tabs. Under Visualisations, several versions of the entire knowledge network are presented displaying selected layers of information (e.g. terms and categories, terms and geni, terms and relations). The Publications tab lists the complete bibliography of project-related articles.

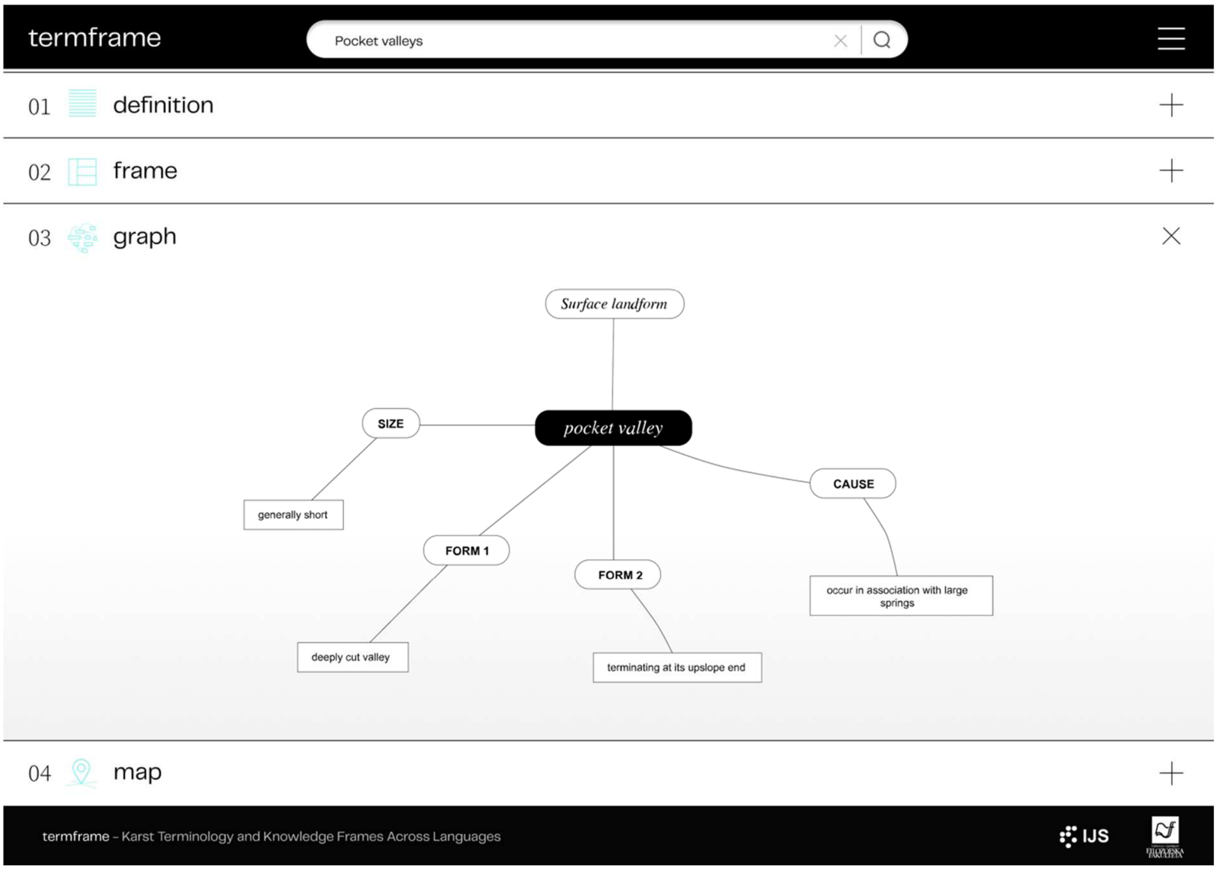


Figure 5: Graph

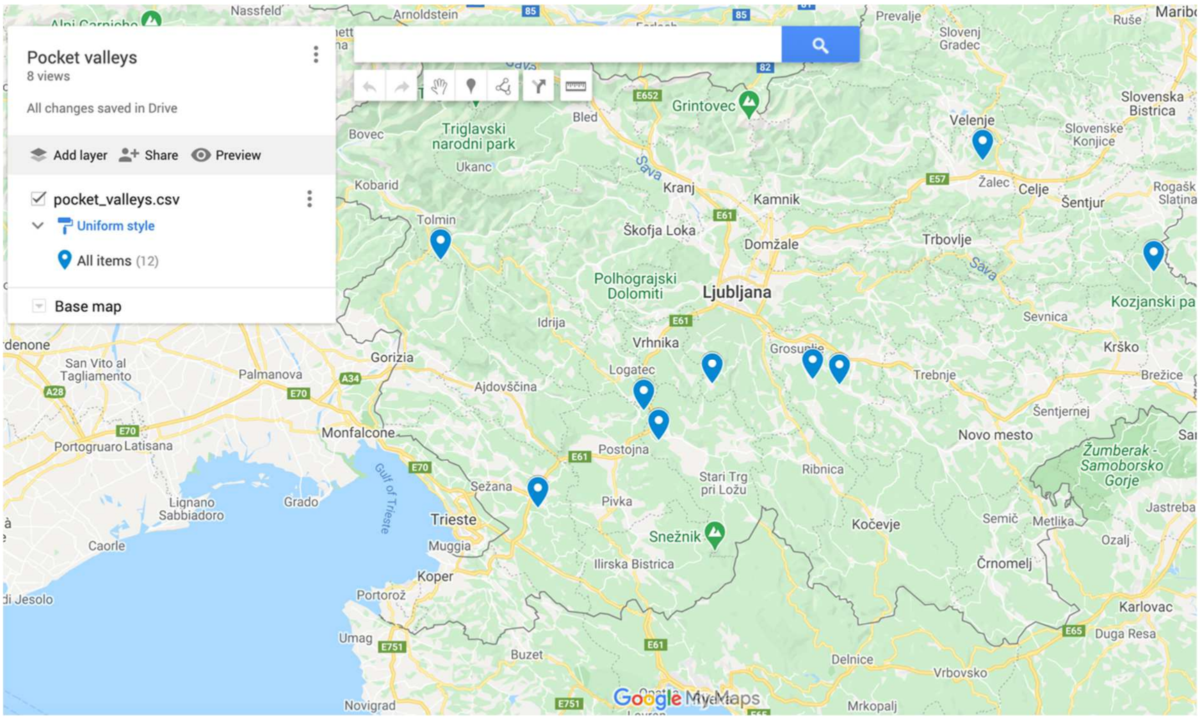


Figure 6: Map of pocket valleys in Slovenia

5. Conclusion

We have described a new resource for karstology which presents structured knowledge in an attractive and innovative manner. The rationale of the design is that even highly specialised knowledge which has partly been obtained using complex text mining techniques can still be accessible and visually compelling. The frame-based restructuring of definitions seems a promising approach which links the textual level of knowledge with the cognitive, spatial and visual spheres.

Since the user interface is still being completed at the time of writing, no usability studies have been performed yet. An evaluation of the web interface by different target groups remains one of our goals for the future. Since karst phenomena in Slovenia and Croatia, but also elsewhere, attract large numbers of visitors who may be interested to explore the karst knowledge base on a hand-held device, we envisage the development of an app which would incorporate location data to the display of maps and images.

Upon project completion (by the end of 2021), several datasets will be made available through the Clarin.si repository for English, Slovenian and Croatian: 1. the TermFrame corpora (except for the works for which distribution was explicitly denied), 2. the extracted and semantically annotated definitions, and 3. the parsed annotations in table format which can be used for visualisation or other form of analysis. The online knowledge base described above is available online without registration.

6. Acknowledgments

This work is supported by the Slovenian Research Agency grant J6-9372 TermFrame: Terminology and knowledge frames across languages, 2018-2021, research programme grant P6-0215 Slovene Language - Basic, Contrastive, and Applied Studies (2019-2024) and by European Union's Horizon 2020 research and innovation programme under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

7. References

Bastian, M., Heiman, S. & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

Bihua, Q. I. U. (2020). A Frame-based Version of NATO Glossaries. China Terminology, 22(3), p. 33.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A. & Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In Proceedings of the LT4DH workshop at COLING 2016, Osaka, Japan.

Faber, P. (2009). The Cognitive Shift in Terminology and Specialized Translation. MonTI. Monografías de Traducción e Interpretación 1, pp. 107-134. https://doi.org/10.6035/MonTI.2009.1.5

Faber, P., León-Araúz, P. & Reimerink, A. (2011). Knowledge representation in EcoLexicon. *Technological innovation in the teaching and processing of LSPs: proceedings of TISLID* 10, pp. 367-386.

Faber, P., ed. (2012). A Cognitive Linguistics View of Terminology and Specialized Language. Berlin/Boston: De Gruyter Mouton.

Faber, P., León-Araúz, P., & Reimerink, A. (2016). EcoLexicon: new features and challenges. GLOBALEX, pp. 73-80.

Giacomini, L. (2018). Frame-based Lexicography: Presenting Multiword Terms in a Technical E-dictionary. In *Proceedings of the XVIII EURALEX International Congress.*

Hick, W.E. (1952). On the rate of gain of information. Quarterly Journal of Experimental Psychology. 4 (4:1), pp. 11–26. doi:10.1080/17470215208416600

León-Araúz, P., Reimerink, A. & Faber, P. (2019). EcoLexicon and by-products: Integrating and reusing terminological resources. Terminology, 25 (2), pp. 222-258.

Mayer T., Terhall, A. & Urban, M. (2014). An Interactive Visualization of Crosslinguistic Colexification Patterns. In Proceedings of VisLR: Visualization as Added Value in the Development, Use and Evaluation of Language Resources, pp. 1-8.

Nielsen, J. (1996). Usability metrics: tracking interface improvements. In IEEE

Software, vol. 13, no. 6, pp. 1-2, Nov. 1996, doi: 10.1109/MS.1996.8740869.

Podpečan, V., Ramšak, Ž., Gruden, K., Toivonen, H. & Lavrač, N. (2019). Interactive exploration of heterogeneous biological networks with Biomine Explorer. Bioinformatics, 24 June 2019, pii: btz509, doi: 10.1093/bioinformatics/btz509.

Pollak, S., Vavpetič, A., Kranjc, J., Lavrač, N. & Vintar, Š. (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. Vienna: KONVENS 2012, pp. 53-60.

Pollak, S., Podpečan, V., Miljkovic, D., Stepišnik, U. & Vintar, Š. (2020). The NetViz terminology visualization tool and the use cases in karstology domain modeling. In Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020), pp. 55–61.

Roche, C., Costa, R., Carvalho, S. & Almeida, B. (2019). Knowledge-based terminological e-dictionaries: The EndoTerm and al-Andalus Pottery projects. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *25*(2), pp. 259-290.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research.13(11), pp. 2498–504.

Stepišnik, U. (2020). Fizična geografija krasa. Ljubljana: Znanstvena založba Filozofske fakultete.

Vintar, Š., Saksida, A., Vrtovec, K. & Stepišnik, U. (2019). Modelling specialized knowledge with conceptual frames: The TermFrame approach to a structured visual domain representation. In I. Kosem et al. (eds.) Proceedings of eLex 2019, pp. 305-318.

Vintar, Š. & Stepišnik, U. (2020). TermFrame: A Systematic Approach to Karst Terminology. Dela, (54), pp. 149-167. https://doi.org/10.4312/dela.54.149-167

Zhou, G. & Xia, J. (2018). OmicsNet - a web-based tool for creation and visual analysis of biological networks in 3D space. Nucleic Acids Research (doi:10.1093/nar/gky510).