Proceedings of the Workshop on Multimodal Wordnets (MMWN-2020), pages 25–32
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Wordnet As a Backbone of Domain and Application Conceptualizations in Systems with Multimodal Data

Jacek Marciniak
Department of Artificial Intelligence, Faculty of Mathematics and Computer Science,
Adam Mickiewicz University in Poznań
Uniwersytetu Poznańskiego 4 Street, 61-614 Poznań, Poland
jacekmar@amu.edu.pl

Abstract

Information systems gathering big amounts of resources growing with time containing distinct modalities (text, audio, video, images, GIS) and aggregating content in various ways (modular e-learning modules, Web systems presenting cultural artefacts) require tools supporting content description. The subject of the description may be the topic and the characteristics of the content expressed by sets of attributes. To describe such resources one can just use some of existing indexing languages like thesauri, classification systems, domain and upper ontologies, terminologies or dictionaries. When appropriate language does not exist, it is necessary to build a new system, which will have to serve both experts who describe resources and non-experts who search through them. The solution presented in this paper used to resource description, allows experts to freely select words and expressions, which are organized in hierarchies of various nature, including that of domain and application character. This is based on the wordnet structure, which introduces a clear order for each of these groups due to its lexical nature. The paper presents two systems where such approach was applied: the E-archaeology.org e-learning content repository in which domain knowledge was integrated to describe content topics and the Hatch system gathering multimodal information about the archaeological site targeted at a wide audience, where application conceptualization was applied to describe the content by a set of attributes.

Keywords: domain and application conceptulizations, wordnet based ontologies, multi-relational and multi-hierarchical indexing languages

1. Introduction

Before building an information system, it is necessary to make a decision regarding the way of organizing the information within so that the data supply process is simple and secure. In order to make the process run smoothly, it is necessary to select the solutions which will support describing objects of a similar kind in a consistent way. This is especially essential when data are input into the system by multiple users working in different time, because there is a risk of describing the same objects in many different ways. Moreover, during the data input various errors will appear, e.g. duplicated entries, incomplete or inconsistent data. In business systems this problem is noticed because of the big scale of this issue. Some researches show that nearly 40% of all company data is found to be inaccurate, or that for instance 92% of businesses admit their contact data is not accurate (Halo, 2020). This creates a need of data cleaning, which takes the form of standardization (replacing of different instances of the same value with one value) or deduplication (detection of duplicate values and their consolidation). These problems appear even when processing data as obvious as e.g. the recipient's address. Therefore, handling them will be a much greater challenge in the case of less obvious data like a type and a nature of pattern of a painting found at an archaeological site (i.e. zoomorfic, geometric, bucranium, wall painting). In such cases, data cleaning must be carried out by experts, who due to little amount of time and working in the project rigor will rarely be available when the data coherence processes will be necessary.

Carrying the data cleaning processes out is always laborious and costly, so it is a wise idea to care about the data coherence when entering them into the system. In order to do this, existing dictionaries, terminologies, thesauri, classification systems or ontologies may be used. This solution may be useful when building systems which are at the advanced stage of the development cycle and store content of universal or well-developed area. Only in such cases it can be assumed that there exists some available indexing language, which would support describing the content in a homogeneous way. Even then, we cannot be sure that all users will perform the process in the same way. Even when dictionary, thesauri or classification system or ontology are used, users can describe the resources in different ways (Hjørland, 2012). This means that they can describe the same object using different words, terms or classes from classification system. The situation is even more complicated when information system is at the initial stage of development and tools supporting resources description do not exists, or existing indexing languages do not comply with the needs due to e.g. cultural differences or domain conceptualization not concordant with the needs of experts responsible for describing the resources.

The paper presents the solution in which the data description is carried out using the indexing language being built during the process of the multimodal data input. The solution has been chosen due to the fact that prior to building two given information systems there was no dictionary, thesaurus, classification system or ontology which would be applicable in the resource description process. In the adopted approach, the experts who input multimodal content into the system, describe it at the same time using freely chosen words or expressions. They are organized into hierarchies and connected with the relations of different nature, including that of domain and application type. The solution is based on the wordnet structure and uses its hierarchy as the core organization of the developed indexing language. Two information systems in which this approach was used are: the E-archeology.org e-learning content repository, where the content description is carried out using the lexical units taken from a wordnet and extended with a domain conceptualization, and the Hatch system storing multimodal data from archaeological site in Çatalhöyük. In the last case, the wordnet structure was supplemented with an application conceptualization.

2. Other solutions for data organisation

In information systems, data structures are in most cases the integral part of the system. An attribute-value approach is used, attributes are part of a system architecture and their organization is determined by programmers during system implementation. Data input by users are added to a database and they can be maintained in it. Other architectures, such as ontology-driven software architectures, allow modelling data structures outside the system (Pan et al., 2013). Such approaches allows to improve the exchange, maintenance and hierarchization of attributes and values assigned to them.

The common programmers' practice is validating data that are input to the system to avoid errors. In the simplest case, validation takes a form of checking the user input with the data type required for the attribute. The input may also be compared with internal dictionary entries. This solution is insufficient when the dictionary can be expanded by users inputting data, or when dictionaries from different systems need to be used. In such cases unexpected errors may appear, such as multiple entries describing the same concept or repeated values.

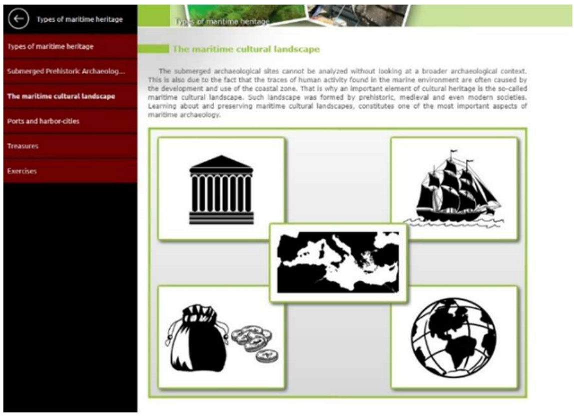
The solution to these problems is using existing dictionaries, terminologies, thesauri, classification systems, ontologies etc. when describing content (Crofts et al., 2010), (Gemet, 2020), (Getty AAT, 2020), (Geonames, 2020), (Iconclass, 2020), (Niles, Pease, 2001). The use of exiting indexing language supports describing the content in a homogeneous way by multiple users. Yet, it forces indexers to refer to the existing conceptualization of the domain, which is why sometimes it may occur impossible to describe the content in a satisfactory way. Dissatisfaction may result from missing terms, hierarchization incompliant with expectations, granularity of concepts and habits of experts describing contents. If an existing controlled vocabulary or classification system is used, and there will be a need to change or add new descriptors to the existing language when indexing, the extension process may be excessively lengthy (Weda, 2016). At times, if the used language is developed by another team, the extension will not be possible at all. Among the problems with using controlled vocabulary to index the resources, there are also: difficulties in differentiating specific and general vocabulary, arbitrariness when defining synonymy and introducing abbreviations or acronyms to vocabulary, adding qualifiers when handling homographs, homonyms, different approach when introducing common and technical terms (Joudrey et al., 2018). Therefore, while dealing with content description, it is beneficial to use a language which allows for maintenance of different types of conceptualization, including the ones that can be extended during the description process and ones that cannot due to their controlled character.

Even if during the description process we use the existing indexing language such as dictionary, thesaurus or classification system, we must remember that access to indexed resources does not necessarily have to be easier (Maniez, 1997), (Hjørland, 2012). It means that during indexing resources stored in some repository, expert responsible for indexing will make arbitrary decisions regarding the use of a particular indexing language. Then, there is a possibility that when describing a concept, one will use more general terms despite the occurrence in a given language of specific terms that allow describing the subject in more detail way. Therefore, there is a need for solution which allows to detect such practices easily and to make corrections without the risk of generating additional errors.

Among numerous approaches to resource indexing, there are some in which a wordnet was used. Princeton WordNet was used, e.g., in indexing the works of arts as complement to other three description systems: Getty AAT, Iconclass and ULAN (Holing et al. 2003). In the LT4EL project, a wordnet was used to index e-learning content stored in LMS Ilias system (Monachesi et al., 2008). The relations used in the solutions were wordnet hyperonymy relation, some relations from Dolce ontology and others from a domain ontology. Some works were also conducted towards mapping thesauri onto wordnets (Maziarz, Piasecki, 2018). plWordNet was also used to enrich a keywords database of the Polish Classification of Activities indexing language (Jastrząb, Kwiatkowski, 2019). In all of these solutions, existing wordnets and other indexing languages were used. Thus, indexers taking part in the content description process could only use descriptors available within those systems.

3. Wordnet enhanced by a domain conceptualization for indexing and searching repository of eLearning content

For the needs of describing the subject of e-learning content stored in the E-archeology.org repository, it was necessary to develop a solution which would allow organizing words and expressions used in the process of resource tagging in a way that supports indexing processes and searching through resources. The repository contains e-learning materials on the protection of archaeological heritage, the management and protection of cultural and natural heritage and introductory materials on archaeology for engineers and engineering for archaeologists (Marciniak, 2014). Currently, the repository contains more than 6,200 learning objects in 9 languages, which together create around 1,700 modules and units, and more than 30 training curricula (Marciniak, 2019a). The content includes text materials, graphics, films, quizzes and animations (Fig. 1).


Types of maritime heritage
Types of maritime heritage
Submerged Prehistoric Archaeolog...
The maritime cultural landscape
Ports and harbor-cities
Treasures
Exercises
Types of maritime heritage

The maritime cultural landscape

The submerged archaeological sites cannot be analyzed without looking at a broader archaeological context. This is also due to the fact that the traces of human activity found in the marine environment are often caused by the development and use of the coastal zone. That is why an important element of cultural heritage is the so-called maritime cultural landscape. Such landscape was formed by prehistoric, medieval and even modern societies. Learning about and preserving maritime cultural landscapes, constitutes one of the most important aspects of maritime archaeology.

Figure 1: E-learning materials in the E-archaeology.org content repository

The e-learning content stored in the repository is compositional and constructed in such a way that allows creating new training curricula from existing modules and units. Initially, the repository contained content regarding protection of archaeological heritage (Marciniak, 2014), and later the materials about management and protection of cultural and natural heritage were added (Marciniak, 2019a).

Considering the large volume of the repository and the diversity of subjects, a proper description of the content is necessary in order to effectively search for modules and units, when new training curricula are compiled. The subject of contents was described by tagging (Smith, 2008). In this process words and expressions freely chosen by an indexer were stored in metadata assigned to e-learning components (keyword metadata from the IEEE LOM scheme). As in the tagging process, the indexers are not limited in terms of tags they use. It is necessary to organize them so that they can be re-used by other taggers. This will allow indexers to choose words and expressions of the appropriate level of detail when the system will propose more than one candidate to choose from.

In order to organize concepts by referring to the knowledge available only to experts, a conceptualization of archaeological and natural heritage domain was introduced into the created indexing language. The wordnet relations in it between words and expressions are intended to provide synonymy support and to allow a distinction of description detail (especially by use of hyperonymy relation) which will both be understandable for experts and non-experts.

3.1 The structure of expanded wordnet and its role in indexing and searching the repository

Words and expressions used during tagging e-learning content were then used to create the PMAH (Protection and Management of Archaeological Heritage) indexing language. At the initial stage of its development finished in 2015, it contained only words and expressions in English and it covered the domain of management and protection of archaeological heritage (Marciniak, 2016). Afterwards, along with providing the repository with a new content, the domain was expanded with management and protection of cultural and natural heritage. It was done by adding new words and expressions and a new domain hierarchy of concepts.

Among words and expressions used when tagging resources, we can distinguish common names (e.g. anthropology, aircraft, aerial archaeology), proper names (e.g. British Museum, Altamira), surnames (e.g. Eric Hobsbawm), geographical names (e.g. Gzira Stadium, France, Europe) and dates (e.g. 1956, 1940–1945).

For the purposes of facilitating the content tagging and searching process by recommendation of more tag candidates to system users (Fig. 2), the words and expressions were connected by the following relations:

– synset to consider the words or expressions as synonymous,

– wordnet relations between synsets (hyperonymy, holonymy, belongs to class),

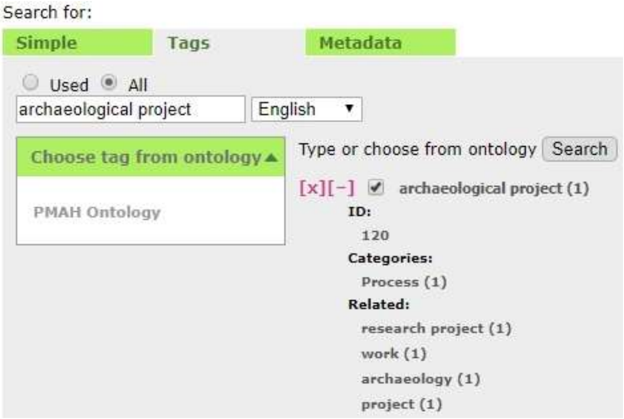
– domain relations between synsets introduced by domain experts,

– generated relations between synsets determining similarity and relatedness of concepts,

– synsets assignment to domain categories determining the domain hierarchy.

The task of wordnet relations is to organize words and expressions in a way that is understandable to all repository users, not only to domain experts. Lexical relations are understandable for all users — both experts and non-experts. Connecting entries using wordnet relations is intended to help the users who do not know the specialized terminology to select of the most appropriate tags when indexing and searching resources. When tagging resources by referring to relations such as hyperonymy / hyponymy (e.g. archaeology – aerial archaeology), holonymy / meronymy (e.g. cultural heritage – cultural heritage management), instance/class (e.g. Altamira - cave), experts can select the tags of an adequate level of detail, increasing the chance of using the tags previously used by other users. The synonymy relation is indicated as one of the basic types of relations used in indexing languages and appears, e.g., in the specification defining the thesauri form (Dextre Clarke, Lei Zeng, 2012). In contrast to controlled vocabularies such as thesauri, the use of synsets to describe synonymy makes indicating the descriptor, i.e. the preferred term impossible. In case of the approach in which indexing of resources takes the form of tagging, this is the expected characteristic of chosen solution. Currently, the words and expressions are grouped in c. 2000 synsets in the PMAH indexing language.

Domain relations between synsets were introduced by the domain experts in order to express the relations of an indefinite nature (e.g. archaeology – archaeological project, heritage – archaeological heritage protection). In the case of PMAH, the used relation was *link*. This relation refers to the fuzzynymy relation from wordnets (Vossen, 2002), (Maziarz et al., 2011) and associative relations from thesauri (Dextre Clarke, Lei Zeng, 2012). Introducing such relations is to allow indexers to access words and expressions connected within the domain. The set of relations between synsets was complemented with the relations defining similarity and relatedness of concepts, which are generated using heuristic rules (Marciniak, 2016). The rules refer to, *inter alia*, wordnet hierarchy (e.g. HasSameHypernym) and produce new relations between synsets to increase the number of tag candidates proposed by the system during tagging and searching through the repository (Fig. 2).



Search for:

○ Used ● All

archaeological project English ▾

Type or choose from ontology

[x][-] ☑ archaeological project (1)

ID:

120

Categories:

Process (1)

Related:

research project (1)

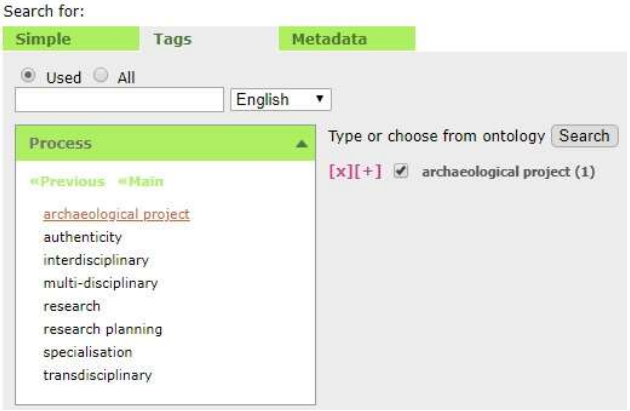
work (1)

archaeology (1)

project (1)

Figure 2: Recommended tag candidates during tagging and searching through the repository

In addition to relations between synsets, all synsets were also mapped onto hierarchical structures created using so-called domain categories (DC). In the adopted approach, the domain categories perform the function of semantic labels used to represent the concepts derived from thesauri, classification systems or domain ontologies. They perform a function analogical to semantic domain from WordNet or domain labels from EuroWordnet allowing a proper organization (categorization) of synsets and being used to group synsets of one semantic field (Fellbaum, 1998), (Vossen, 2002). The hierarchical structure of domain categories is created using the generic or mereological relations. The conceptualization obtained by means of domain categories hierarchy and synsets mapped onto them, is of the multi-facated character. Initially, 12 the most general domain categories were placed at the top of the domain categories hierarchy. When words and expressions from new subject domains were used as tags during uploading the contents of different subject into the repository, the number of domain categories at the top of the hierarchy increased to 26. Among them, there are categories like Archaeological heritage, Archaeological process, Chronology, Archaeology, Landscape, Nature, Policy, etc. Now, the number of all domain categories is 238. The hierarchy of domain categories with assigned synsets is presented to users searching and tagging the repository as a hierarchical index (Fig. 3).


Search for:

Simple | Tags | Metadata

● Used ○ All

English ▾

Process ▴

«Previous «Main

archaeological project
authenticity
interdisciplinary
multi-disciplinary
research
research planning
specialisation
transdisciplinary

Type or choose from ontology Search

[x][+] ☑ archaeological project (1)

Figure 3: Index of tags in the domain categories hierarchy

The structure of the wordnet hierarchy extended with domain relations and domain categories hierarchy is presented in Fig. 4. The indexing language created in such a way can be considered as an ontology understood as an arrangement of objects appearing in a given domain and the knowledge about them shared by specialists or as a specification of conceptualization (Joudrey et al., 2018), (Gruber, 1993). The wordnet based ontology thus understood, following Uschold's and Grunninger's (1996) formalization, will be considered as a semiformal ontology.

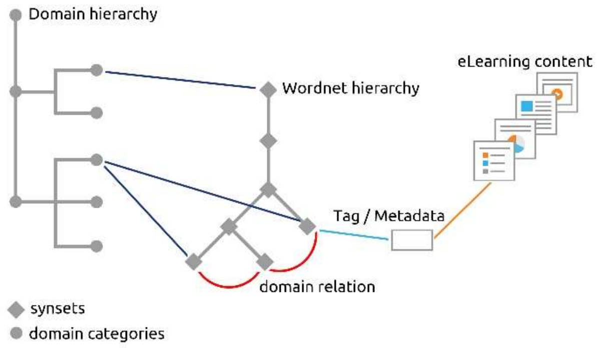
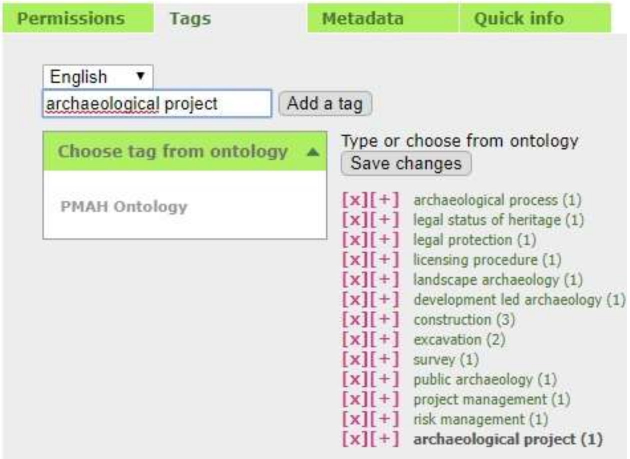


Figure 4: The structure of the wordnet hierarchy extended with domain relations and domain categories hierarchy

3.2 Building extended wordnet while tagging eLearning resources

The PMAH ontology was developed along with the expansion of content in the E-archaeology.org repository. At the early stage of the development, the initial set of words and expressions (c. 1,000) used by experts in the process of content tagging was than expanded by additional 400 entries (synonyms, more general terms and terms connected with associative relations) (Marciniak, 2016). The hierarchical structure for these entries was developed on the basis of the existing wordnet structure (i.e. Princeton WordNet) considered as a referential wordnet according to the algorithm of wordnet based ontology creation (Marciniak, 2016). In the case of the PMAH ontology, the algorithm aimed to integrate all words and expressions used by taggers into ontology. It expanded the ontology only in those fragments in which a new synset was included. It did not aimed to incorporate all synsets from the referential wordnet, only hyperonyms of the new introduced synset were added. According to the algorithm, domain relations (i.e. associative relations) between synsets were added by domain experts. They also created the hierarchy od domain categories and mapped synsets onto them.

At the second stage of development, when the repository was expanded with the content from management and protection of cultural and natural heritage domain, additional 600 words and expressions were added into the PMAH ontology. At this stage, the process of adding all new words or expressions to the ontology took place directly during tagging e-learning materials. Because the ontology was already built and contained a substantial set of entries, the system suggested to an indexer words or expressions used earlier in the repository as tags by other indexers (Fig. 5).


Permissions | Tags | Metadata | Quick info

English ▾

archaeological project | Add a tag

Choose tag from ontology ▴

PMAH Ontology

Type or choose from ontology

Save changes

[x][+] archaeological process (1)
[x][+] legal status of heritage (1)
[x][+] legal protection (1)
[x][+] licensing procedure (1)
[x][+] landscape archaeology (1)
[x][+] development led archaeology (1)
[x][+] construction (3)
[x][+] excavation (2)
[x][+] survey (1)
[x][+] public archaeology (1)
[x][+] project management (1)
[x][+] risk management (1)
[x][+] archaeological project (1)

Figure 5: Tagging resources in the repository

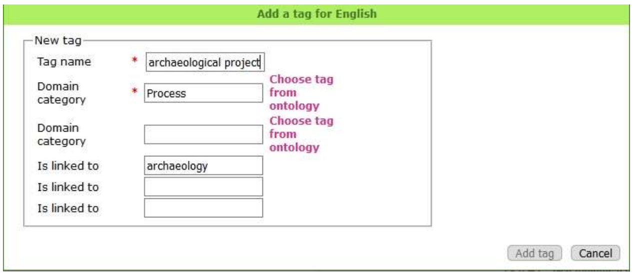
If the indexer (an expert), did not found a candidate to be used as a tag among words and expressions from the ontology, he or she always could add a new tag. Such a tag was assigned to e-learning content metadata and added at the same time to the PMAH ontology. This process was performed in two steps:

– the expert's task was to assign the word or expressions to one or multiple domain categories, add synonyms or associative relations with other synsets from the ontology,

– a lexicographer added afterwards the unit to the wordnet hierarchy.

The first actions were undertaken in the content repository at the time of content tagging with the use of one combine form (Fig. 6). The expert could choose domain categories onto which the introduced tag had to be assigned to, as well as word or expression from the ontology to be connected with the associative relations.

The latter actions were performed outside the e-learning content repository in a dedicated tool - the Ontology Repository Tool (Marciniak, 2019b). Using the external tool allowed the introduction of necessary modifications and extensions into the ontology, such as typo corrections, removal of duplicates or hierarchy adjustments. It allowed to carry out ontology maintenance processes by knowledge engineers (domain experts) who did not needed to be supported by programming teams.



Add a tag for English

New tag

Tag name * archaeological project

Domain category * Process

Domain category

Is linked to archaeology

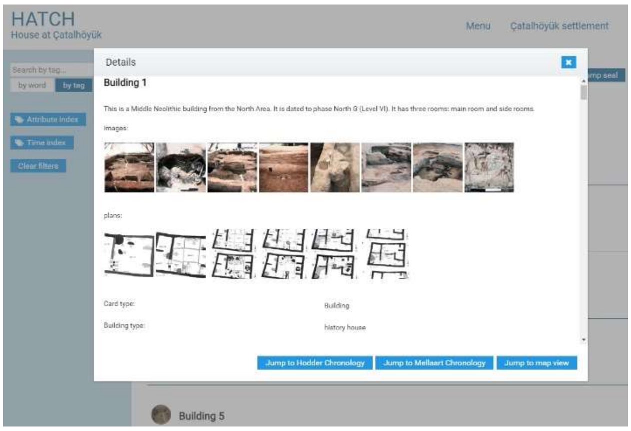
Is linked to

Is linked to

Figure 6: Adding an expression to the domain structure of the PMAH ontology and linking it to a synset

4. Wordnet enhanced by an application conceptualization for describing artefacts of heterogenous character

The second application of the solution based on wordnet structure enhanced by the expert knowledge, was the use of the extended PMAH ontology in the Hatch system. The Hatch (House at Çatalhöyük) is an advanced Web system designed to create and maintain a digital collection (Marciniak et. al, 2020). The Hatch is aimed at presenting a wide range of multimodal data about the Neolithic settlement at Çatalhöyük in a multiscalar and interactive form. It combines information of different character (types of artefacts, their attributes, relations among them) with different form of their presentation (text, photographs, graphics, maps, GIS localizations and multiscalar chronology of artefacts). It is designed to meet the needs and expectations of both professionals and general public interested in the human past (Fig. 7).



Details

Building 1

This is a Middle Neolithic building from the North Area. It is dated to phase North G (Level VI). It has three rooms: main room and side rooms.

images:

plans:

Card type: Building

Building type: history house

Figure 7: A card of an artefact in the Hatch system

The system was constructed when the excavation works at the site were very advanced. At the time of the system construction, the expert team already had a large amount of various data about the site, such as text descriptions, photographic material, maps, GIS database and artefacts chronology. Yet, the data were not organized in terms of their presentation in a system for users without specialist knowledge about Çatalhöyük site. Due to the character of the archaeological site, there was no indexing language which could be used to describe the resources stored in the Hatch system. Therefore, a solution was adopted in which the PMAH ontology was extended with entries related to the Neolithic site, with consideration to the character of Çatalhöyük. Furthermore, the Hatch system is to be supplemented with e-learning courses, which will supply the E-archaeology.org repository and will have to be tagged in a similar way to other resources stored there.

In contrast to tagging e-learning content in the repository, where all words and expressions chosen as tags by indexers are assigned to one metadata, entries from the PMAH ontology are assigned as a values to multiple attributes describing artefacts in the Hatch system. The number of attribute organization schemes equals the number of object types stored in the Hatch. Their arrangement results from the need to present the data in the system and that is why it has simply applicational character.

4.1 The structure of extended wordnet and its role in describing artefacts of different type

Words and expressions which extended the PMAH ontology were obtained during data input into the Hatch system. The artefacts are organized in the system, in so-called cards, where attribute-value structures serve to describe artefacts' characteristics . The number of attribute-value pairs is different for each object type and the corresponding card. For instance, the attributes for an Imagery card are *Imagery type* and *Motifs*, respectively taking exemplary values of *wall painting* and *zoomorfic*. In Animal bones card for *Animal bones types* attribute, the exemplary values are *astragali*, *crane ulna* or *scapula*.

Attribute-value structures were constructed using a new domain category type and the synsets assigned to them. The new domain category (DC-HATC) is different than the one used in the previously presented solution used for tagging e-learning content, because the character of a new hierarchical arrangement of concepts in the PMAH ontology built to accomplish the Hatch system needs, is also different. The approach in which attribute-value structures are built with domain categories embedded in the ontology, make possible the storage and maintenance of the data outside the Hatch system. It implements the postulate of getting the information structures out from the information system, which streamlines the process of correcting words and expressions used for resource indexing.

The fact that information structures are hosted outside the information system facilitates the use of the same word or expression as values assigned to several attributes. This creates a possibility to reuse the word or expression which was used before as the value in a different attribute. For example, the *zoomorfic* value was used as a value of two attributes describing the motif type: in Stamp seal card and Imagery card. Thanks to this, the user searching through the system will receive the cards of two different types when typing the *zoomorfic* value as the query to the system.

Due to the character of the archaeological artefacts for which words and expressions were used as attributes' values, the words and expressions used in the Hatch system can be divided into:

– units equivalents of which can be found in the largest reference wordnets, e.g. plWordnet (plWordnet, 2020) or Princeton WordNet (WordNet Search, 2020), e.g. bucranium, flint, geometric, kerb, relief,

– units for which equivalents could not be found in any referential wordnet, including units which could be and those which couldn't be added there for different reasons e.g. astragali, animal bone, crane ulna, abandonment deposit, zoomorfic,

– units with a strong terminological character, e.g. barley seeds, feasting deposit, post retrieval pit, multi-roomed construction,

– expression referring to the time, e.g. "3–12 years – child", "20+ adult",

– chronology in qualitative units (TP M, Level II, North I).

For the purposes of the Hatch system, the words and expressions were connected by the following relations:

– synsets connecting words and expressions considered to be synonymous,

– wordnet relations between synsets (hyperonymy, holonymy, belongs to class),

– generated relations between synsets determining similarity of concepts,

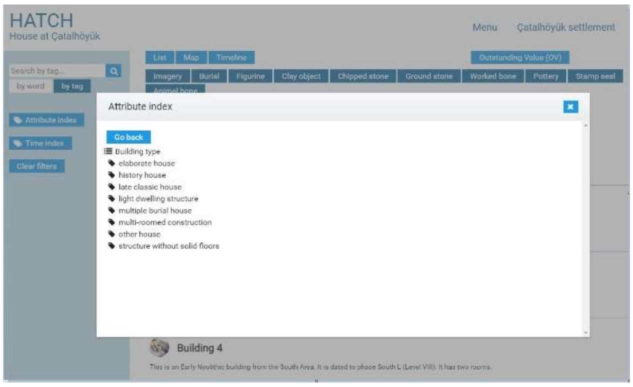
– synsets assigned to domain categories,

– special relations for handling object dating,

In the Hatch system, synonymy relation was used to keep the information about the singular or plural form of words or expressions used as descriptors. There is no general rule regarding the use of singular or plural in descriptors. It depends on the specific language and the regulations adopted by the individual community or country (Joudrey et al., 2018). In the Hatch system, in a situation where singular and plural was used as values of the same attribute, this was not considered as an error and was not corrected. Instead, both forms were related in one synset. The solution is not canonical and was adopted because of the practical matters. In the process of synsets creation, an interesting problem of ambiguity arose. For instance, in the case of a word building it was necessary to make a decision whether it fulfils the definition from the referential wordnet, or it is necessary to introduce a new meaning and create a new synset due to the character of the Neolithic buildings located at the Çatalhöyük site. The first solution was chosen, despite it may be debatable in the case of domain and applicational uses of the PMAH ontology.

Wordnet relations were used to relate those words and expressions (synsets) which were found in the referential wordnet, as well as for those which could not be found. This approach was adapted due to the need of the rules generating relations between synsets determining concepts similarity, which use lexical relations, especially hyperonymy relation. As in the case of the e-learning content repository, the generated relations determining similarity and relatedness between synsets are intended to be used in recommendation of best tag candidates to the Hatch non-expert users searching the system.

Similarly to the e-learning content repository, all synsets were mapped onto hierarchical structures built using domain categories. Due to a different character of this hierarchy, other type of category was used (DC-HATC). This hierarchical arrangement of words and expressions is useful only in the case of the Hatch system because of its strongly applicational character. At the top of the domain categories hierarchy, there are three categories which arrange words and expressions considering their role in the Hatch system: Attributes, Auxiliary attributes and Time Index. Other domain categories being attributes of cards (Animal bones, Figurine, Imagery, Pottery, etc.) are subcategories of the Attributes category. In general, there are 57 domain categories arranging words and expressions taking the Hatch needs into account. Domain categories hierarchy with assigned synsets is presented to the users searching through the Hatch system as two separate hierarchical indexes: Attribute index (Fig. 8) and Time index (Fig. 9).

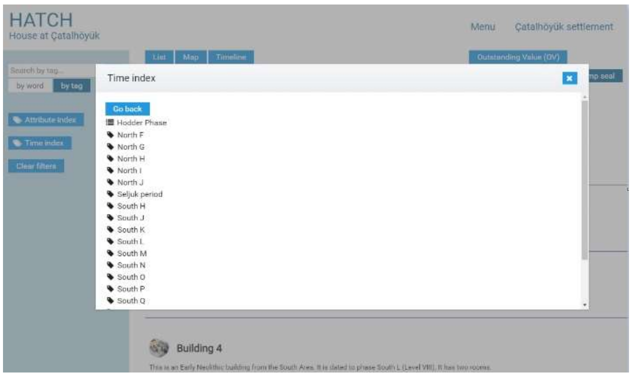


Building type
- elaborate house
- history house
- late classic house
- light dwelling structure
- multiple burial house
- multi-roomed construction
- other house
- structure without solid floors

Building 4

This is an Early Neolithic building from the South Area. It is dated to phase South L (Level VIII). It has two rooms.

Figure 8: Attribute index in the Hatch system



Hodder Phase
- North F
- North G
- North H
- North I
- North J
- Seljuk period
- South H
- South J
- South K
- South L
- South M
- South N
- South O
- South P
- South Q

Building 4

This is an Early Neolithic building from the South Area. It is dated to phase South L (Level VIII). It has two rooms.

Figure 9: Time index in the Hatch system

Time index shows a special use of relations from the PMAH ontology for chronology arrangement of artefacts at the Çatalhöyük site. Due to the character of the site, chronological order is arranged with qualitative values. Absolute dating using C14 method is available only for selected objects. Therefore, when presenting the chronology of the objects in the Hatch system, three different systems developed for the needs of Çatalhöyük site were used: Mellaart Phase, Hodder Phase and TP Phase. Each system consists of a set of highly terminological values, e.g. North F, Level III, TP M. As the timeline with artefacts from the site is one of the ways of presenting the objects in the Hatch system, it was necessary to assign qualitative values used in the chronology system to particular dates, so that the date can be interpreted in a programming component used to create the timeline. As in the case of other values assigned to attributes, terms from a chronology system (e.g. North F) are also assigned to domain categories from the PMAH ontology. Those entries were connected with dates (e.g. 6300 BC) specifying the approximate and conventional (from the point of view of the archaeological research methodology) time of a given period. The relations used have associative and applicational character, i.e. they are not useful outside the Hatch system.

The structure of the wordnet hierarchy extended with domain and applicational hierarchies is presented in Fig. 10.

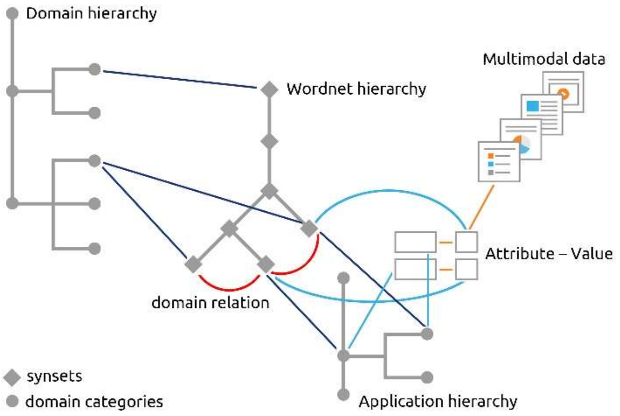


Figure 10: The structure of the wordnet hierarchy extended with domain and application hierarchies

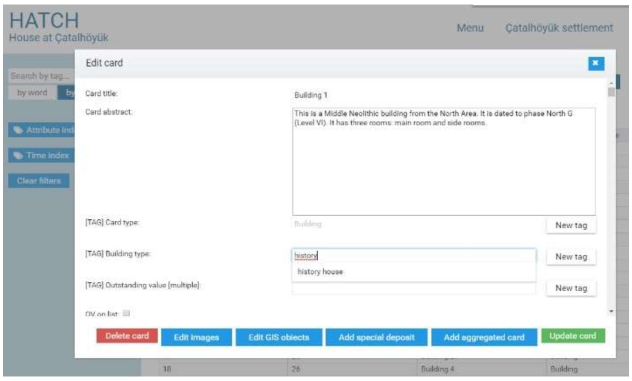
4.2 Building extended wordnet while describing multimodal data

The PMAH ontology was extended for the needs of organizing the information in the Hatch system when the system was fulfilled with the multimodal data such as photos, maps, GIS data, text descriptions and bibliographic references. They were grouped into cards corresponding to different artefacts types,. In total, 725 cards, 1107 photos, 194 maps and 71 000 GIS objects were input into the system. The process of supplying the system with the data was carried out by a few domain experts for about a year. For the domain experts (archaeologists), it was mainly the task of ordering information about artefacts from the site, choosing appropriate photographic materials, locating the object on the GIS map and determining the chronology of artefacts. Assigning words or expressions as values of attributes was performed simultaneously to other actions and was not prominent. As there was a risk of errors appearing in the process, values were assigned to attributes in one form directly in the Hatch system. Its goal was to minimalize the number of errors appearing when several experts were extending the PMAH ontology at the same time. The goal was achieved when assigning values to attributes due to (Fig. 11):

– suggesting by the system words or expressions which were used before by other indexers as a value in an attribute,

– suggesting by the system words or expressions which were not used before as a value in the attribute, but which were present in the ontology due to the fact that they were either assigned as a value to another attribute before, or were just present in the ontology, but not yet used in the Hatch system,

– entering new words or expressions and assigning them as a value to a particular attribute.



Card title: Building 1

Card abstract: This is a Middle Neolithic building from the North Area. It is dated to phase North G (Level VI). It has three rooms: main room and side rooms.

[TAG] Card type: Building

[TAG] Building type: history

history house

[TAG] Outstanding value [multiple]:

OV on list: □

Figure 11: Adding a word to the domain structure of the PMAH ontology during entering data in the Hatch system

Words and expressions entered during artefacts description by domain experts, as well as the PMAH ontology, were placed in the external tool, which was used for maintenance tasks. The maintenance of the data was periodically handled by one domain expert, who controlled the entered words and expressions and introduced corrections such as deleting the entries with errors (e.g. typographic errors), deleting the values inconsistent with the description criteria adapted by the team, replacing too general or too detailed values and the ones of an inappropriate granularity. Other deleted elements included incorrect values resulting from the software engineering errors and internet connection errors.

5. Conclusion

The solution presented in this paper shows that in the process of indexing resources of different character and highly specialized subject, it is necessary to use indexing languages which allow to extend them according to the needs with maintaining the clear organization of terms at the same time. Application of wordnet based ontology using the wordnet structure as a backbone of the whole system, allows to use arrangement resulting from the wordnet and refers to conceptualization available for both experts and non-experts. Due to such structure, a non-expert will be able to switch between specialized terminology and words and expressions known from common language, thanks to the tag candidates recommendation facility available in the presented systems. This will allow non-experts to formulate more appropriate queries when searching through the repository. Experts will be able to choose the most appropriate level of detail when indexing a resource. Incorporation of domain and applicational conceptualizations to the system allows distinguishing different arrangement of terms meeting different needs in one indexing language. Domain ordering allows experts to arrange entries according to their specific needs and knowledge. Applicational ordering improves the process of resource description, as it allows using words and expressions already used before for indexing resources by other experts.

Due to the separation of knowledge structures outside the system in which they are used, it is possible to carry out ontology maintenance processes by knowledge engineers who do not need to be supported by programming teams. This makes the ontology maintenance process more clear and keeps the indexing consistent, when the action is performed by multiple users. This creates a possibility to introduce changes and extensions to the indexing language without changing the IT structure of the system in which this indexing language is used.

6. Bibliographical References

Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M. (ed.) (2010). Definition of the CIDOC Conceptual Reference Model, ICOM/CIDOC.

Dextre Clarke, S.G., Lei Zeng, M. (2012). From ISO 2788 to ISO 25964. The Evolution of Thesaurus Standards towards Interoperability and Data Modeling, ISO Information Standards Quarterly, Winter 2012, vol. 24.

Fellbaum, Ch. (ed.) (1998). WordNet: An Electronic Lexical Database, MIT Press.

Getty AAT. (2014). About the AAT, . Access data: February 2020.

Gemet. (2012). Gemet: General Multilingual Environmental Thesaurus, http://www.eionet.europa.eu/gemet/en/about/, Access Date: February 2020.

Geonames. (2020): The GeoNames geographical database, http://www.geonames.org, Access date: February 2020.

Halo. (2020). Data Quality in BI the Costs and Benefits, https://halobi.com/blog/infographic-data-quality-in-bi-the-costs-and-benefits/, Access data: February 2020.

Iconclass RKD (2020): IconClass, www.iconclass.nl, Access Date: February 2020.

Hjørland, B. (2012). Is classification necessary after Google?, Journal of Documentation, vol. 68, iss. 3, pp. 299-317.

Hollink, L., Schreiber, G., Wielemaker, J., Wielinga, B. (2003). Semantic annotation of image collections. S. Handschuh, M. Koivunen, R. Dieng, S. Staab (eds.), Proceedings of the KCAP'03 Workshop on Knowledge Capture and Semantic Annotation, Florida, October 2003, s. 41-48.

Jastrząb, T., Kwiatkowski, G. (2019). Enriching a Keywords Database Using Wordnets – a Case Study. Proceedings of the 10th Global Wordnet Conference, Wrocław, July 23-27 2019, pp. 329-335.

Joudrey, D.N, Taylor, A.G., Wisser, K.M. (2018): The Organization of Information, 4th ed. Libraries Unlimited.

Maniez, J. (1997). Database merging and the compatibility of indexing languages, Knowledge Organization, vol. 24, no. 4, s. 213-224.

Marciniak, A., Marciniak, J., Filipowicz, P., Harabasz, K., Hordecki, J. (2020). Engaging with the Çatalhöyük database. House at Çatalhöyük (HATCH) and other applications. Near Eastern Archaeology, 83:2.

Marciniak, J. (2014). Building E-learning Content Repositories to Support Content Reusability, In International Journal of Emerging Technologies in Learning (iJET), Volume 9, Issue 3 (2014), pp. 45-52.

Marciniak J. (2016). Building wordnet based ontologies with expert knowledge. Zygmunt Vetulani, Hans Uszkoreit, Marek Kubis (ed.) Human Language Technology. Challenges for Computer Science and Linguistics Papers, Lecture Notes in Computer Science, Vol. 9561, pp. 243-254, Springer International Publishing.

Marciniak J. (2019a). Methods and Tools for Centers of Integrated Teaching Excellence Providing Training in Complementary Fields. Proceedings of 11th International Conference on Computer Supported Education (CSEDU 2019) - Volume 2, pp. 527-534.

Marciniak J. (2019b). Ontology Repository Tool for effective development and deployment of wordnet based ontologies. Proceedings of 9th Language and Technology Conference (LTC 2019), Human Language Technologies as a Challenge for Computer Science and Linguistics - 2019, pp. 25-26.

Maziarz, M., Piasecki, M., Szpakowicz, S., Rabiega-Wiśniewska, J. (2011). Semantic relations among nouns in Polish Wordnet grounded in lexicographic and semantic tradition, Cognitive Studies, vol. 11.

Maziarz, M., Piasecki, M. (2018). Towards Mapping Thesauri onto plWordNet. Proceedings of the 9th Global Wordnet Conference, Singapore, 8-12 January 2018, pp. 45-53.

Monachesi, P., Simov, K., Mossel, E., Osenova, P., Lemnitzer, L. (2008). What ontologies can do for eLearning. Proceedings of IMCL 2008. 16-18 April 2008.

Niles, I., Pease, A. (2001). Towards a Standard Upper Ontology, Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS'01) – Volume 2001, October 2001, ACM, pp. 2-9.

Pan, J., Staab, S., Aßmann, U., Ebert, J., Zhao, Y. (eds.) (2013). Ontology-Driven Software Development. Springer, 2013

plWordnet (2020): plWordnet, Słowosieć, a large network http://plwordnet.pwr.wroc.pl/wordnet/, Access Date: February 2020.

Uschold, M., Gruninger, M. (1996). Ontologies: principles, methods, applications, Knowledge Engineering Review, vol. 11(2), s. 93-136.

Weda, R. (2016). Update on the Dutch AAT work. https://www.getty.edu/research/tools/vocabularies/weda_zelfde_aat_dutch_2016.pdf, Access : February 2020.

WordNet Search (2020). Wordnet Search 3.1. http://wordnetweb.princeton.edu/perl/webwn. Access : February 2020.

Vossen, P. (ed.) (2002). Euro WordNet General Document. Version 3, University of Amsterdam.