Terminology Dictionary Digitalization

Volodymyr Shyrokov [1], Iryna Ostapova[1], Yevhen Kupriianov[2], Alona Dorozhynska[1], Mykyta Yablochkov[1] and Iuliia Verbynenko[1]

[1] *Ukrainian Lingua-Information Fund of NAS of Ukraine,3, Holosiivskyi avenue, Kyiv, 03039, Ukraine*
[2] *National Technical University "Kharkiv Polytechnic Institute", Kyrpychova str. 2, Kharkiv, 61002, Ukraine*

Abstract

One of the most important tasks of Ukrainian lexicography is the elaboration of technology for conversion of the whole dictionary heritage into digital format. Many national dictionaries that have been traditionally published in paper, are being digitalized now due to the current world trends. This purpose requires elaboration of adequate technological solution. In this context it should be noted that there have been elaborated various approaches to dictionary digitalization. However, a general solution to be applicable for dictionaries of different types hasn't been found yet. Therefore, the purpose for our research is to build up and propose digitalization technology which would be common and usable for different dictionaries. The Dictionary of Ukrainian Biological Terminology, which has a rather large volume and complex structure, was chosen for digitalization. Our technology proposed represents the step-by-step conversion of the dictionary from paper text to web site version. The basic steps are as follows: 1) text in PDF-format, 2) HTML-file, 3) formal model of the dictionary referred to as lexicographic system, 4) XML-file, 5) database, 6) website. The first step was converting the PDF to a simple HTML file that contains only visual markup. The next, but main stage, was developing the model of the dictionary lexicographic system to serve as a basis for the XML-structure of the dictionary entry. The further digitalization was based on the XML file. The dictionary text was marked up with XML tags using special software. At the next steps the database and website were elaborated. With the website interface the user has not only the access for updating and revision of the dictionary text but the every-time technical support.

Keywords

Computer lexicography, lexicographic system, parsing, XML, database, digital space, website.

1. Introduction

One of the tasks resolved by modern computer lexicography is creating digital dictionaries, in particular multilingual terminology dictionaries. Most of them don't have digital versions, so the urgent task is their digitalization. Many tools which have been created today are applicable only for individual stages of terminology dictionary making process however there is no universal technological solutions to the basic problems of digital terminography. This is especially true of the digital reception of traditional terminological heritage, especially multilingual [5,6]. Among all the dictionary diversity, the Dictionary of Ukrainian Biological Terminology was chosen for digitization [1] (according to the authors, this dictionary is the first lexicographical work of the new generation in Ukrainian studies, covering the most common biological terminology in Ukrainian, Russian and English and offering term definitions). The proposed terminology dictionary covers the normative general scientific and widely

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland
EMAIL: vshirokov48@gmail.com (V. Shyrokov); irinaostapova@gmail.com (I. Ostapova); eugeniokuprianov@gmail.com (Y. Kupriianov); alonochkatkachyk@gmail.com (A. Dorozhynska); gezartos@gmail.com (M. Yablochkov); yulia_verbinenko@yahoo.co.uk (I. Verbynenko)
ORCID: 0000-0001-5563-8907 (V. Shyrokov); 0000-0001-8221-3277 (I. Ostapova); 0000-0002-0801-1789 (Y. Kupriianov); 0000-0001-6554-6731 (A. Dorozhynska); 0000-0003-1175-1603 (M. Yablochkov); 0000-0002-7111-0755 (I. Verbynenko)

used terminology of biological sciences, fixed in modern encyclopedic, general and special dictionaries, as well as in scientific, popular science, educational and informative literature.

Our approach offers stage-by-stage conversion of the dictionary text into website format. The main stages are as follows: Paper dictionary => Lexicographic system (L-system) of the dictionary => XML-tagging of the dictionary text following the L-system structure => Converting XML-tagged text of the dictionary into database format => Web site version of the dictionary on ULP (Ukrainian linguistic portal). This technological procedure, in our opinion, contains the steps which are possible to be applied to other dictionaries, so we believe that this sequence is an effective and universal way to transform paper dictionaries into digital format.

2. Method

2.1. Term dictionary conceptual model

The digital transformation of lexicographic works requires some general theoretical framework to describe and represent the widest possible range of the objects in lexicography. Our developments are based on the theory of lexicographic systems. The dictionary is considered as an information system of a special type, namely as a lexicographic one. This is an abstract language-information object focused on the implementation of a comprehensive description of the lexical and grammatical structures of the language or a set of languages [3, 4]. The system architecture corresponds to the standard three-level architecture of information systems ANSI/X3/SPARK, according to which the information system is divided into conceptual, internal and external data levels [3]. The internal level defines the types, structures and formats in which data are to be represented, stored and manipulated. The external level ensures a set of procedures which allow the user to manipulate the data represented at the internal level. The conceptual level of representation (conceptual model) is a symbolic, semantic model which integrates the various specialists' views about the domain in an unambiguous, final and inconsistent way.

As a conceptual model we have chosen the lexicographic data model [3] which is represented in a simplified form:

$$\{D,\ I_Q^0(D), V(I_Q^0(D)), \beta, \sigma[\beta], Red[V(I_Q^0(D)]\}, \quad (1)$$

where D is the modeling object (domain), in our case the Dictionary of Ukrainian Biological Terminology; $I_0(D) = \{x_i\}$ is the set of the language units described in the dictionary (in the theory of lexicographic systems it is usually referred to as the *set of elementary information units*); $V(I_0(D))$ is a set of descriptions (interpretations) of elementary information units (in case of dictionaries the set $V(I_0(D)) = \{V(x_i)\}$ corresponds to the set of dictionary entries dedicated to words x_i); β indicates a set of structural elements to be revealed in the process of the dictionary text analysis; $\sigma[\beta]$ designates the structure to be generated within β by the operator σ and represents the system of relationships reflecting the semantics of the domain considered; the restriction of $\sigma[\beta]$ by $V(x)$ gives the microstructure $\sigma(x)$ of the dictionary entry $V(x)$; $Red[V(I_0(D)]$ is the mechanism of recursive reduction which reveals the finest structures of the lexicographic system. The structures β and $\sigma[\beta]$ specify the semantics of linguistic facts and regularities composing the lexicographic system (L-system). In this case β is a set of the simplest information elements of the dictionary (words, abbreviations, entry notes, numbers, elements of grammar and vocabulary description etc.). The structures β are given explicitly in the dictionary and defined in the following way: a set of $\beta(x)$ forming the entry $V(x)$ is assigned to each $x \in I^Q(D)$ so that:

1. $x \in \beta(x)$
2. Any fragment of the dictionary entry $V(x)$ can be built from the elements belonging to $\beta(x)$
3. The principle of identifying and defining of each $\beta(x)$ is to be common for all $V(x)$ with headwords $x \in I^Q(D)$

2.2. Dictionary entry structure

The conceptual model has been built taken into account the paper version of the dictionary in question. That is, the typographic design, layout and structure of printed texts of dictionary articles are analyzed, which are interpreted as identifiers of the corresponding elements of lexicographic structures β and σ[β]. The following elements compose the dictionary entry structure:

- **СС**: dictionary entry (represented by the paragraph in the text)
- **ЗТ_У**: head term in Ukrainian (НО means homonym number as an attribute of the head term)
- **ТБ$_i$**: term block (a text line composed of Ukrainian, Russian and English terms as well as their parameters)
- **СМБ$_i$**: explanatory block (entire dictionary entry text without ТБ), the number of explanatory blocks corresponding to that of term blocks

To clarify the dictionary entry structure, we have divided each term block into various sub-blocks which are separated from each other by the language marker:

- **ПТБ_У**: Ukrainian sub-block (includes the whole text of the dictionary entry in Ukrainian together with all the parameters)
- **ПТБ_Р**: Russian sub-block (includes the whole text of the dictionary entry in Russian together with all the parameters)
- **ПТБ_А**: English sub-block (includes the whole text of the dictionary entry in English together with all the parameters)

In its turn a sub-block may comprise several complexes dedicated to Ukrainian, Russian and English term:

- **ТК_У$_{1\text{-}n}$:** complex for Ukrainian term (each group includes one term and all its parameters in Ukrainian)
- **ТК_Р$_{1\text{-}n}$:** complex for Russian term (each group includes one term and all its parameters in Russian)
- **ТК_А$_{1\text{-}n}$:** complex for English term (each group includes one term and all its parameters in Russian)

The term complexes are comprised by: language marker (ММ), grammar note before the term (ГРД), explanatory note (СР), term (Т) and grammar note after the term (ГРП). There can be one term in one complex. The language marker has been introduced in the complex for Ukrainian term for generalization. The structure of the complexes for Ukrainian, Russian and English terms is represented below:

ТК_У$_{1\text{-}n}$	Complex for Ukrainian term
ММ	language marker (*укр.*) introduced for generalization
ГРД	grammar note before the term (all grammar parameters placed before Ukrainian term)
Т_У	Ukrainian term
ГРП	Grammar note after the term (all grammar parameters placed after Ukrainian term)
ТК_Р$_{1\text{-}n}$	Complex for Russian term
ММ	marker for the Russian language (identified as *рос.* in the text)
СР	explanatory note (all explanatory parameters)
ГРД	Grammar note to the term
Т_Р	Russian term
ГРП	Grammar note after the term
ТК_А$_{1\text{-}n}$	Complex for English term
ММ	language marker (identified as *англ.* in the text)
СР	explanatory note
ГРД	Grammar note to the term
Т_А	English term
ГРП	Grammar note after the term

The explanatory block (СМБ) is constituted by the definition blocks (БТ$_m$). There are as many explanatory blocks as there are definitions in the dictionary entry. Each explanatory block consists of:

- **НТ:** definition number
- **ТЛ:** definition
- **СРТ:** explanatory note to ТЛ

- **ПБТС:** collocation sub-block including all the collocations composed by the term
- **БП:** reference block
- **БСН_ЗТ:** block of synonyms to the head term

The collocation sub-block is made up the blocks of term collocations (there may be several blocks). The term block of collocations consists of a term block of a collocation (**ТБСЛ**), explanatory block of collocations (**БТСЛ**) and block of synonyms to collocations (**БСН_СЛ**). Similarly, we introduce sub-blocks of the term block of collocations.

The sub-blocks are introduced for the Ukrainian, Russian and English languages: ПТБСЛ_У, ПТБСЛ_Р, ПТБСЛ_А. Each of them may include several complexes for term collocations. The complexes are formed by the language marker (ММ), grammar note before the term collocation (ГРСД), term collocation (ТС) and grammar note after the term collocation. Explanatory notes before the term collocation haven't been revealed. There can be only one term collocation in a complex.

ПТБСЛ_У term collocation sub-block for Ukrainian
ПТБСЛ_Р term collocation sub-block for Russian
ПТБСЛ_А term collocation sub-block for
ТСК_У$_{1\text{-}р}$ sub-block of the term block for Ukrainian
ТСК_Р$_{1\text{-}р}$ sub-block of the term block for Russian
ТСК_А$_{1\text{-}р}$ sub-block of the term block for English
ММ language marker
ГРСД grammar note before the term collocation
ГРСП grammar note after the term collocation
ТС_У term collocation in Ukrainian
ТС_Р term collocation in Russian
ТС_А term collocation in English

Let us consider the explanatory note for term collocations (БТСЛ) consisting of the definition number of the term collocation (НТСЛ) and definition itself (ТЛС). The synonym blocks to the head term (БСН_ЗТ) and to the term collocation (БСН_СЛ) consist of the synonym marker (МС) and synonym row ($СН_1 \ldots СН_n$):

- **МС:** synonym marker ("Син.")
- **$СН_1 \ldots СН_n$:** synonym row (to be included both by terms and term collocations)

The reference block (БП) consists of the sub-blocks (ПБП). Each one includes the whole array of the references. The sub-blocks are subdivided into the number of references. The number of sub-blocks corresponds to that of reference markers (МП). The reference marker is the note "*див.*" (see). The sub-blocks are also subdivided into the addressee (**САНТ**) and recipients (**$САТ_j$**). The general XML scheme to mark-up the dictionary is as follows:

```
<СС> Entry
    <ЗТ_У homonym number=i>Ukrainian head term</ЗТ_У>
    <ТБ number=i> Term block
        <ТК_У number=i> Ukr. term complex
            <Т_У> Ukrainian term</Т_У>
            <ГРД number =i> Grammar note to</ГРД>
            <ГРП number =i> Grammar note to</ГРП>
            <ММ> укр.</ММ>
        </ТК_У>
        <ТК_Р number = i> Rus. term complex
            <Т_Р> Russian term </Т_Р>
            <СР> Explanatory note</СР>
            <ГРД number =i> Grammar note to</ГРД>
            <ГРП number =i> Grammar note to</ГРП>
            <ММ> рос.</ММ>
        </ТК_Р>
        <ТК_А number =i> Engl. term complex
            <Т_А> English term</Т_А>
            <СР> Explanatory note</СР>
```

```
                    <ГРД number=i> Grammar note to</ГРД>
                    <ГРП number=i> Grammar note to</ГРП>
                    <ММ> англ.</ММ>
                </ТК_А >
        </ТБ>
        <СМБ number =i >
                <БТ number =i> Explanatory block
                    <ТЛ> Definition </ТЛ>
                    <НТ> Definition </НТ>
                    <СРТ> Explanatory note </СРТ>
                    <СИН_ЗТ number=i> Synonym block
                                <Т_У> term</Т_У>
                                <ТС_У> term</ТС_У>
                                <МС> Син. </МС>
                    </СИН_ЗТ>
                    <БТС number=i> Term collocations block
                            <ТБСЛ number =i> Term collocation block
                                <ТКС_У number =i> Ukrainian term collocation complex
                                        <ТС_У> Term collocation_</ТС_У>
                                        <ГРС> Grammar note</ГРС>
                                        <ММ> Language marker</ММ>
                                </ТКС_У>
                                <ТКС_Р number =i> Russian term collocation complex
                                        <ТС_Р> Term collocation</ТС_Р>
                                        <ГРС> Grammar note </ГРС>
                                        <ММ> Language marker </ММ>
                                </ТКС_Р>
                                <ТКС_А number =i> English term collocation complex
                                        <ТС_А> Term collocation </ТС_А>
                                        <ГРС> Grammar note </ГРС>
                                        <ММ> Language marker</ММ>
                                </ТКС_А>
                            </ТБСЛ>
                            <БТСЛ number =i> Term collocation explanatory block
                                <ТЛС> Definition to collocation </ТЛС>
                                <НТЛС> Number of definition collocation</НТЛС>
                            </БТСЛ>
                            <СИН_СЛ number=i> Synonym block
                                <Т_У> term</Т_У>
                                <ТС_У> term</ТС_У>
                                <МС> Син. </МС>
                            </СИН_СЛ >
                    </БТС>
                    <БП number = i> Reference block
                            <ПБП number = i> Reference sub-block
                                <САНТ> addressee</САНТ>
                                <САТ number=i> recipient </САТ>
                                <МП> reference marker <МП>
                            </ПБП>
                    </БП>
                </БТ>
        </СМБ>
</СС>
```

2.3. Example of marking the dictionary entry with XML tag

The example below shows the printed version of the entry arrangement which corresponds to the developed entry structure in XML format.

двохо#дкові, -их, *ім*., *мн*. (*рос*. двухо#дковые, *англ*. ringed lizards, worm lizard) 1. Червоподібні плазуни, тіло яких укрите суцільною роговою плівкою, поділеною на квадрати поздовжніми і поперечними борозенками.

Entry elements:

ТБ [term block]: **двохо#дкові**, -их, *ім*., *мн*. (*рос*. двухо#дковые, *англ*. ringed lizards, worm lizard)

ПТБ_У [Ukrainian sub-block]: **двохо#дкові**, -их, *ім*., *мн*.

ТК_У [Complex for Ukrainian term]: **двохо#дкові**, -их, *ім*., *мн*.

ЗТ [Head term]: **двохо#дкові**

ГРП [Grammar note after the term]: -их, *ім*., *мн*.

ПТБ_Р [Russian sub-block]: *рос*. двухо#дковые

ТК_Р [Complex for Russian term]: *рос*. двухо#дковые

ММ[Language marker]: *рос*.

Т_Р [Russian term]: двухо#дковые

ПТБ_А [English sub-block]: *англ*. ringed lizards, worm lizard

ТК_$А_1$ [Complex for English term]: *англ*. ringed lizards

ТК_$А_2$ [Complex for English term]: worm lizard

ММ [Language marker]: *англ*.

Т_$А_1$ [English term]: ringed lizards

Т_$А_2$ [English term]: worm lizard

СМБ [Explanatory block]: 1. Червоподібні плазуни, тіло яких укрите суцільною роговою плівкою, поділеною на квадрати поздовжніми і поперечними борозенками.

БТ [Definition block]: 1. Червоподібні плазуни, тіло яких укрите суцільною роговою плівкою, поділеною на квадрати поздовжніми і поперечними борозенками.

НТ [Definition number]: 1

ТЛ [Definition]: Червоподібні плазуни, тіло яких укрите суцільною роговою плівкою, поділеною на квадрати поздовжніми і поперечними борозенками.

The XML text reflecting the entry structure of the term dictionary in question is as follows:

```
<СС>
<текст_СС><![CDATA[<B>двохо#дкові</B>, -их, <I>ім.</I>, <I>мн.</I> (<I>рос.</I> двухо#дковые, <I>англ.</I> ringed lizards, worm lizard) червоподібні плазуни, тіло яких укрите суцільною роговою плівкою, поділеною на квадрати поздовжніми і поперечними борозенками.]]></текст_СС>
<ЗТ homonymy number='0'>двохо#дкові</ЗТ>
<ТБ number="1">
    <ТК_У number="1">
        <Т_У>двохо#дкові</Т_У>
        <ГРП>-их, ім., мн.</ГРП>
        <ММ>укр.</ММ>
    </ТК_У>
    <ТК_Р number="1">
        <Т_Р>двухо#дковые</Т_Р>
        <ММ>рос.</ММ>
    </ТК_Р>
    <ТК_А number="1">
        <Т_А>ringed lizards</Т_А>
        <ММ>англ.</ММ>
    </ТК_А>
        <ТК_А number="2">
        <Т_А>worm lizard</Т_А>
        <ММ>англ.</ММ>
```

```
        </ТК_А>
        </ТБ>
        <СМБ number="1">
            <БТ number="1">
            <НТ>1</НТ>
            <ТЛ>червоподібні плазуни, тіло яких укрите суцільною роговою плівкою, поділеною на квадрати поздовжніми і поперечними борозенками</ТЛ>
            </БТ>
        </СМБ>
    </СС>
```

3. Experiment

3.1. Dictionary text representation in lexicographic database structure

The programming language and technological platform for development were chosen, respectively, the C# language and .Net 5 platform. Due to the object structure of dictionary entry representation, there has been used a documentary-type database that meets the following requirements:

- Ease of use
- Possibility of supporting transaction mechanisms
- Possibility of parallel access to database
- Free of charge for research purposes.

As a result, the choice was made for LiteDB (https://www.litedb.org/), a database of documentary type, created as a relatively simple, free copy of the shareware database MongoDB. An additional advantage of this database is the ease of installation and connection of the software package, as LiteDB is implemented as a single library file (dll) and a single configuration file (xml), rather than the entire software package. This database is informally called an analogue of MySQL for documentary databases.

A parsing library Html Agility Pack (https://html–agility–pack.net) was used to process the obtained XML files in the software environment.

For developing the structure of the repository class, two opposing approaches were considered: 1) creating a "family" of classes, where each class was a separate structural element, and the relationship between them is a reference to instances of the respective classes; 2) using nested classes, where the whole hierarchy of structural elements is part of the main parent class. Within goals set, bringing all the structural elements into separate independent collections in the database (which is a direct consequence of the first approach) is too redundant, and the implementation of access to them unreasonably increases the complexity of program logic. The second approach was further modified by converting structural elements from nested classes to nested structures to optimize the continued use of dictionary entry classes by the application.

Each dictionary article is presented in the internal model of the application by the class of 1st type:

- Class "DictionaryStorageClass": container for the dictionary entry decomposed in various structural elements.

So, the application uses a documentary database, the data is stored identically to their representation in the internal model. To ensure the coherence and efficiency of the development process as well as the use of classes-repositories and classes of the of dictionary entries index (described herein after), there have been identified several types of constants:

- Language list: enumerator **Languages**, values **Ukrainian**, **Russian**, **English**.
- List of term structure characteristics: enumerator **TerminologyStructures**, values **Word**, **Collocation**.
- List of term types: enumerator **TerminologyTypes**, values **MainTerm**, **SecondaryTerm**, **LinkedTerm**, **Synonym**.
- List of language markers: array of text variables **LanguageMarks**, values "**укр.**", "**рос.**", "**англ.**".

The class of "**DictionaryStorageClass**" type has the following structure in the lexicographic database:

- Dictionary entry identifier: integer variable **ID**.
- Head term: text variable **OriginalDicEntryString**.
- Homonymy indicator of head term: integer variable **Omonim**.
- Original text of the dictionary entry in text line format: text variable **OriginalDic EntryString**.
- List of term blocks in the dictionary entry: list of elements **TerminologyBlock** – **TermsList**.
- List of explanatory blocks in the dictionary entry: **SemanticBlock** – **Semantic BlocksList**.
- Entry text HTML format, generated on the basis of the class: text variable **Dic EntryHTMLString**.
- Entry text generated on the basis of the class: text variable **DicEntryNoTags String**.

The element of "**structTerminologyBlock**" type (Term block) is represented by the following variables:

- Identifier for implicit connection of the term block with explanatory block: integer variable **LinkingID**
- List of term complexes in the given block: elements list of **TerminologyComplex** type **TerminologyComplicesList**

The element of "**structTerminologyComplex**" type (Term complex) is represented by the following variables:

- Term: text variable **Term**
- Notes followed by the term (explanatory notes): list of text variables **Semantic RemarksList**
- Notes before the term (grammar notes): list of text variables **GrammaticRemarks LeadingList**
- Notes after the term (grammar notes): list of text variables **GrammaticRemarksFollowing List**
- Sequence number for visualization: integer variable **SequenceNumber**
- Language marker for visualization: text variable **LanguageMark**
- Language indicator: variable **Language** of **Languages** type
- Term structure indicator: variable **TerminologyStructure** of **Terminology Structures** types
- Term type indicator: variable **TerminologyType** of **TerminologyTypes** type

The element of "**structSemanticBlock**" type (Explanatory block) is described by the following variables:

- Identifier for implicit connection of explanatory block with term block: integer variable **LinkingID**.
- List of definition blocks of the given explanatory block: elements list **InterpretationsList** of **InterpretationBlock** type.

The element of "struct **InterpretationBlock**" (Definition block) is represented by the following variables:

- Term definition: term variable **Interpretation**
- Identifier for implicit connection of collocation definition block with term collocation block, or sequence number for visualization of definition of the terms of "word" type: integer variable **LinkingID**
- Notes after definition (explanatory notes): list of text variables **SemanticRemarks List**
- List of synonyms: list of text variables **SynonymsList**
- List of references to the definitions in other entries list of variables **LinksList** of **InterpretationLink** type
- List of collocations **CollocationsList** of **CollocationBlock** type

The element of "**structInterpretationLink**" (reference element) is represented by the following variables:

- Reference term: text variable **LinkingTerm**
- Head term in the reference entry: text variable **LinkedDicArticleTerm**
- Homonymy index of head term: integer value **LinkedDicArticleTermOmonim**
- Identifier of reference entry: integer variable **LinkedDicArticleID**

- Analogue of the term referred to in the entry – text variable **ReferenceTerm**
- Text marker of reference element: text variable **LinkTypeMarker**

The element of "**structCollocationBlock**" type (Collocation block) is represented by the following variables:

- Sequence number for visualization: integer variable **SequenceNumber**
- List of term complexes in the given block: elements list **CollocationsTermsList** of **TerminologyComplex** type
- List of definition blocks in the given collocation block: elements list **Collocation InterpretationsList** of **InterpretationBlock** type

A complete diagram of the relationships of the class-repository and its nested structural elements is shown in Figure 1.

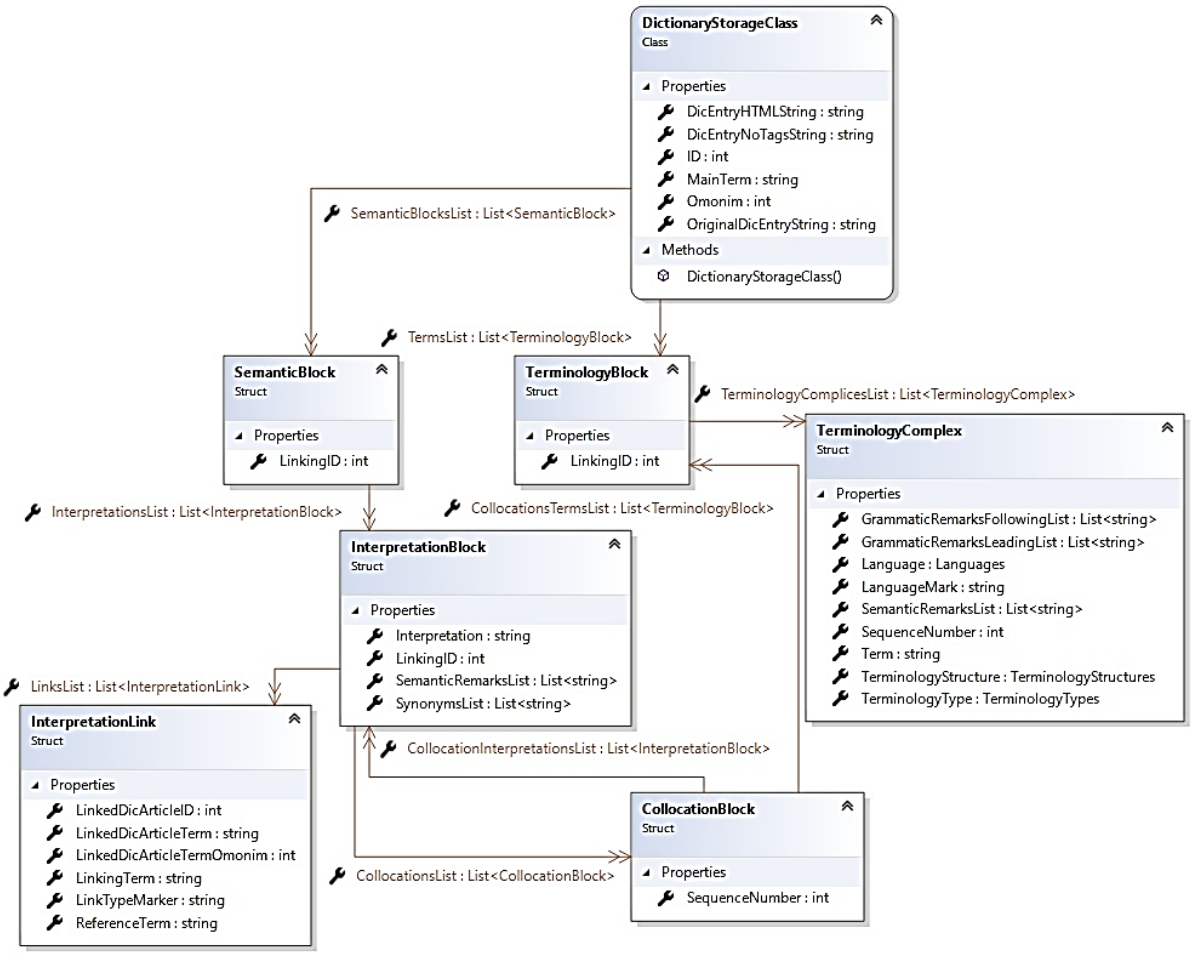

Figure 1: Diagram of the relationships of the class-repository and its nested structural elements

The process of parsing the dictionary entries does not contain any technological features. Owing to the direct relationship between the elements of the XML-structure and the elements of the class-repository, parsing is reduced to "passing" through this structure and filling the corresponding elements of the class.

The only step in parsing, which isn't completely trivial, was the processing the created link elements (**InterpretationLink**). It was performed after processing the input XML file, creating and writing to the database all classes-repositories of the entries to make possible the search for dictionary entries by head term of the referenced entry (elements **LinkedDicArticleTerm** and **LinkedDicArticleTerm Omonim**), and by recording entry ID in the reference element (**LinkedDic ArticleID**). The examples of class-repository **DictionaryStorageClass** are given below.

```
{
 "_id": 1,
 "MainTerm": "двохо#дкові",
 "Omonim": 0,
 "OriginalDicEntryString": "двохо#дкові, –их,
ім., мн. (рос. двухо#дковые, англ. ringed lizards,
worm lizard) червоподібні плазуни, тіло яких
укрите суцільною роговою плівкою,
поділеною на квадрати поздовжніми і
поперечними борозенками.]]>",
 "TermsList":
 [
  {
   "LinkingID": 1,
   "TerminologyComplicesList":
   [
    {
     "Term": "двохо#дкові",
     "GrammaticRemarksFollowingList":
     [
      "–их",
      "ім",
      "мн"
     ],
     "SequenceNumber": 1,
     "LanguageMark": "укр.",
     "Language": "Ukranian",
     "TerminologyStructure": "Word",
     "TerminologyType": "MainTerm"
    },
    {
     "Term": "двухо#дковые",
     "SequenceNumber": 1,
     "LanguageMark": "рос.",
     "Language": "Russian",
     "TerminologyStructure": "Word",
     "TerminologyType": "MainTerm"
    },
    {
     "Term": "ringed lizards",
     "SequenceNumber": 1,
     "LanguageMark": "англ.",
     "Language": "English",
     "TerminologyStructure": "Word",
     "TerminologyType": "MainTerm"
    },
    {
     "Term": "worm lizard",
     "SequenceNumber": 2,
     "LanguageMark": "англ.",
     "Language": "English",
     "TerminologyStructure": "Word",
     "TerminologyType": "SecondaryTerm"
    }
   ]
  }
 ],
 "SemanticBlocksList":
 [
  {
   "LinkingID": 1,
   "InterpretationsList":
   [
    {
     "Interpretation": "червоподібні плазуни,
тіло яких укрите суцільною роговою плівкою,
поділеною на квадрати поздовжніми і
поперечними борозенками",
     "LinkingID": 1
    }
   ]
  }
 ]
}
```


Figure 2: Class-repository **DictionaryStorageClass** for the entry "*двоходкові*"

An integral part of modern digital dictionaries is the extensive indexing of dictionary entry elements. When developing the index class of the dictionary entry, as in the case of the repository class, the development of the dictionary in two stages and laying the groundwork for expanding the list of head words was taken into account. Based on this, the following decisions were made:

- All elements of the index must be functionally equal
- Indexes of word terms bear the information about the relevant explanatory block and the entry in general
- Indexes of term collocations carry the information about the relevant collocation block and the entry in general
- If it is impossible to fill in the information on the relevant structural element of the article to which the index refers, appropriate mark is made

Each index element is represented by the application inner model by the class of the first type:

- Class of "**DictionaryIndexClass**" type: index element container

The class of "**DictionaryIndexClass**" type in lexicographic database has the following structure:

- Inner identifier for DB: integer variable
- Indexed term: text variable **Term**
- Homonymy index for indexed term: integer variable **Omonim**
- Entry identifier: integer variable **DicArtID**
- Language indicator: variable **Language** of **Languages** type
- Term structure indicator: variable **TerminologyStructure** of **Terminology Structures** type
- Term type indicator – variable **TerminologyType** of **TerminologyTypes** type
- Availability indicator of grammar notes for the term: Boolean variable **HasTermGram Remarks**

- Availability indicator of explanatory notes for the term: Boolean variable **HasTermSem Remarks**
- Availability indicator of explanatory notes for at least one definition of the term: Boolean variable **HasInterpSemRemarks**
- Indicator of filling in the information on the relevant structural element of the entry – Boolean variable **InfoFilled**
- Number of term definitions: integer variable **InterprNum**
- Number of term collocations: integer variable **CollocNum**
- Number of synonyms: integer variable **SynonymsNum**
- Number of references in term definitions: integer variable **LinksNum**
- Number of term complexes in the term block: integer variable **TermComplices Num**
- Number of explanatory blocks in the entry: integer variable **ArticleSemBlocks Num**

An instance of this class is created for the following elements of the dictionary entry:

- For each term complexes (words and collocations)
- For all synonyms from definitions and collocations
- For all references from definitions and collocations

For synonyms, the index contains information about the corresponding explanatory block. For references, the index can contain two versions of information:

- If **ReferenceTerm** was found among the created indexes, the information is duplicated from it
- If the element is not found, the information is taken from the index element of the head term of the article

The example in figure 3 below shows the class of the entry index elements **DictionaryIndexClass**.


```
{
 "_id": 4,
 "Term": "worm lizard",
 "Omonim": 0,
 "DicArtID": 1,
 "Structure": "Word",
 "Type": "SecondaryTerm",
 "Language": "English",
 "HasTermGramRemarks": false,
 "HasTermSemRemarks": false,
 "HasInterpSemRemarks": false,
 "InfoFilled": true,
 "InterprNum": 1,
 "CollocNum": 0,
 "SynonymsNum": 0,
 "LinksNum": 0,
 "TermComplicesNum": 4,
 "ArticleSemBlocksNum": 1
}
```


Figure 3: Entry index elements for the head term *worm lizard*

The term index "*worm lizard*" of the entry «*двохо#дкові*» is as follows: Structure – Word; Type – Secondary term; Language – English; Availability of grammar notes to the term – no; Availability of explanatory notes to the term – no; Availability of explanatory notes in the entry – no; Number of definitions – 1; Number of collocations – 0; Number of synonyms – 0; Number of references – 0; Number of term complexes in the block – 4, Number of explanatory blocks – 1.

The interface (external model) of the dictionary incorporates the developments and experience gained in developing the toolkit for researching the Spanish dictionary (Diccionario de la lengua Española 23 ed.) [2], Ukrainian-Polish Lexicon of Active Phraseology and application for visualization of Etymological Dictionary of Ukrainian Language (EDUL) with functions of superficial analysis of the entries and their comparison with the printed version of EDUL [5].

The HTML code for entry visualization is created dynamically for all entries during the application launch and is stored in a temporary collection in the database. Using the capabilities provided by HTML 5 makes possible to enter a large amount of information into the HTML-code of the entries both for visual presentation of articles and to provide interactive functionality (currently – the transition to active parcel elements). The example of the entry "*двохо#дкові*" in HTML format and for user's view is given below.

HTML code:

<article> двохо#дкові 1. <mark class= "LangMark">укр.</mark> <mark class="TermWord" >двохо#дкові</mark>, <mark class="GramRem GramRemFollow">–их, ім, мн</mark>, <mark class="LangMark">рос.</mark> <mark class="TermWord"> двухо#дковые

</mark>, <mark class= "LangMark">англ.</mark> <mark class="TermWord"> ringed lizards </mark>, <mark class="TermWord">worm lizard</mark> червоподібні плазуни, тіло яких укрите суцільною роговою плівкою, поділеною на квадрати поздовжніми і поперечними борозенками; </article>

User's view:

двохо#дкові

1.

укр. двохо#дкові, -их, ім, мн, *рос.* двухо#дковые, *англ.* ringed lizards, worm lizard

червоподібні плазуни, тіло яких укрите суцільною роговою плівкою, поділеною на квадрати поздовжніми і поперечними борозенками;

4. Results

Based on the L-system model and lexicographic database structure the following requirements were set for the dictionary interface:

- Displaying the linear text of the dictionary articles with color highlight of specific structural elements of the entries
- Providing access to all elements of the dictionary wordlist with the ability to use them while searching in the dictionary
- Making possible to make samples conforming the parameters available in the index class (signs of dialectics, onomastics and homonymy)
- Providing the possibility of conducting a full-text search on the content of the dictionary entry
- Ensuring the possibility of navigating by the links from one entry to another with recording the navigation history

For dictionary interface development it was decided to use .Net Core technologies to ensure multi-platform application, and WebAPI to ensure data exchange, namely the processing of queries between the client and server parts of the web application. Since the task was to visualize dictionary entries, not to edit them, the interface was developed in this regard – visualization of the entry, variations of search in the word list and making samples of dictionary entries on the available parameters. For easy creation and further development of the interface elements, a set of HTML, CSS and JavaScript scripts in the Bootstrap language was used, which ensures quick creation and deployment of necessary interface elements. The main interface elements are the word list window and the window for displaying the dictionary entry.

5. Conclusions

The described parsing scheme of Dictionary of Ukrainian Biological Terminology is actually universal and suitable to be used in creating digital versions of almost any three- (and multi-) lingual terminology dictionaries based on their PDF-texts. This versatility is achieved by combining the following factors:

1. Applying the theory of lexicographic systems, which is universal and adequately reflects the structure of dictionaries of any kind. The three-level architecture of the L-system in the form of ANSI / X3 / Spark provides ample opportunities for conceptual generalizations, software modifications, variations of interaction scenarios of different users with the system, etc.
2. Application of methodology and technology of converting digital PDF-text of the dictionary into lexicographic database using the sequence: dictionary text in PDF => dictionary text in Word format => HTML text => XML text.

This approach allows the presentation of complex structured lexicographic information in the form of a well-formed XML document reflecting the hierarchy of the information contained in a typical dictionary entry. This is achieved through the implementation of an abstract lexicographic model that adapts the semantic properties of arbitrary special information. The conversion of XML text to the lexicographic database is performed automatically, which determines the high efficiency of this parsing method.

The availability of dictionary text in XML is a real prerequisite for creating various applications, including virtual systems of professional interaction such as VLL (virtual lexicographic laboratory), modification of source dictionary material, its integration into other dictionaries, use as material for creating resident systems of professional information processing (editing, abstracting, automatic translation, conceptual design and knowledge engineering, etc. [7–12]).

6. References

[1] D. M. Grodzinsky, L. O. Simonenko and other, Ukrainian biological terminology Dictionary, KMM, Kyiv, 2012.

[2] Real Academia Española: Diccionario de la lengua española, 23.ª ed., [versión 23.5 en línea]. URL: https://dle.rae.es.

[3] V. A. Shyrokov (Ed.), Linguistic and information studies: works of the Ukrainian Language and Information Fund NAS of Ukraine, volume 1: Scientific paradigm and basic language and information structures, Ukrainian Lingua-Information Fund of NAS of Ukraine, Kyiv, 2018. URL: https://movoznavstvo.org.ua/files/tom_1_B5_print.pdf. doi: 10.33190/978-966-02-8683-2/8684-9.

[4] V. A. Shyrokov (Ed), Linguistic and information studies: works of the Ukrainian Language and Information Fund NAS of Ukraine, volume 2: Grammar systems, Ukrainian Lingua-Information Fund of NAS of Ukraine, Kyiv, 2018.

[5] I. Kernerman, A multilingual trilogy: Developing three multi-language lexicographic datasets, in: Proceedings of Electronic Lexicography in the 21st Century: Linking lexical data in the digital age, eLex2015, Herstmonceux Castle, United Kingdom, 2015, pp. 372–383p. URL: https://elex.link/elex2015/.

[6] L. Trap-Jensen, Lexicography between NLP and linguistics: aspect of theory and practice. In: J. Čibej, V. Gorjanc, I. Kosem, S. Krek (eds.), Lexicography in Global Contexts, Proceedings of the 18th EURALEX International Congress, Ljubljana, 2018, pp. 25–38.

[7] L. Giacomini, Frame-based Lexicography: Presenting Multiword Terms in a Technical E-dictionary, in: Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana University Press, Faculty of Arts, Ljubljana, 2018. URL: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1.pdf

[8] M. Czerepowicka, The structure of a dictionary entry and grammatical properties of multi-word units, in: Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference, Lexical Computing CZ, Brno, 2021. URL: https://elex.link/elex2021/wp-content/uploads/eLex_2021-proceedings.pdf.

[9] T. Mészáros, M. Kiss, The DHmine Dictionary Work-flow: Creating a knowledge-based author's dictionary, in: Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana University Press, Faculty of Arts, Ljubljana, 2018. pp. 77–86. URL: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1.pdf.

[10] P. Storjohann, Commonly confused words in contrastive and dynamic dictionary entries, in: Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana University Press, Faculty of Arts, Ljubljana, 2018, pp. 187–197. URL: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1.pdf.

[11] E. Sassolini, A. F. Khan, M. Biffi, M. Monachini, S.Montemagni, Converting and Structuring a Digital Historical Dictionary of Italian: A Case Study, in: Electronic lexicography in the 21st century: smart lexicography. Proceedings of the eLex 2019 conference, Sintra, 2019, pp. 603–621. URL https://doi.org/10.5281/zenodo.3726847.

[12] J. Norri, M. Junkkari, T. Poranen, Digitization of data for a historical medical dictionary, Lang Resources & Evaluation 54 (2020) 615–643. URL: https://doi.org/10.1007/s10579-019-09468-2.