Baltic J. Modern Computing, Vol. 11 (2023), No. 1, 150-160
https://doi.org/10.22364/bjmc.2023.11.1.09

NLP-PIPE Incorporation into ONTO6 Framework Workflow

Uldis STRAUJUMS

University of Latvia, Faculty of Computing, Raina bulvaris 19, Riga, LV-1586, Latvia

uldis.straujums@lu.lv

Abstract. The article proposes an approach to simplify a process of identifying significant concepts for a given domain. The author describes his ONTO6 methodology based on a semi-informal meta-ontology. The stages of applying the ONTO6 methodology are: the development of a meta-ontology instance appropriate for the domain to be informatized; the development of an initial ontology from the meta-ontology instance; and the gradual detailing of domain concepts that appear in the initial ontology – the development of an enriched initial ontology. The transition of ONTO6 methodology to ONTO6 framework by usage of tools – NLP-PIPE, Python custom programs, Protégé plugin Cellfie – is demonstrated.

Keywords. ontology learning, domain-specific modeling, NLP-PIPE

1. Introduction

The article continues author's approaches to learn significant concepts for a given domain. The article is an extended version of the author's article (Straujums, 2022). The author has done research in the specific field of informatization (Straujums, 2008; Straujums, 2010). Informatization is understood as an analysis of the business processes, the specification of requirements and the development of software.

In his previous research the author had proposed a unified description of methods and suggestions to identify the essential concepts of the domain to be informatized, to introduce notations for the various levels of detail, and to specify details for the informatization aspects. The author's approach helped overcome the difficulties observed during implementation of several informatization projects and to implement several improvements:

- The development of a unified understanding about the domain to be informatized, particularly about the essential concepts and their interpretation
- The introduction of a suitable notation for various aspects of informatization which are necessary for users involved in the project and appropriate for different levels of competence
- The proposal of a general methodology for performing informatization.

The original ONTO6 methodology has been applied to informatization-specific domains. The ONTO6 methodology appears to be expandable to other domains with several enhancements.

The goal of the paper is to show how to transform the original informatization-specific ONTO6 methodology into the ONTO6 framework suitable for the essential concept elicitation for a given domain.

The essential concept elicitation is language specific because the role of a concept has to be figured out mainly from text sentences. Therefore an automatic text analysis by NLP tools is of a great value for the ONTO6 methodology.

The structure of the rest of the paper is – a description of related works, a description of the ONTO6 methodology, a description of steps needed to transform the ONTO6 methodology into the ONTO6 framework, a proof-of-concept demonstration of a basic ONTO6 framework workflow.

2. Related work

Methodologies in the development of ontologies reflect the formal background of the ontology developer.

There are several approaches known for ontology development, for example:

- **Logical theories**. The specification can be developed in the form of a logical theory that describes the intended meaning. Nicola Guarino (Guarino, 1998) implements this principle, "An ontology is a logical theory accounting for the intended meaning of a formal vocabulary, i.e. its ontological commitment to a particular conceptualization of the world. The intended models of a logical language using such a vocabulary are constrained by its ontological commitment. An ontology indirectly reflects this commitment (and the underlying conceptualization) by approximating these intended models."
- **Linguistic relativism.** A specification can be developed using the concept of linguistic relativism, an approach suggested by Boris Wyssusek (Wyssusek, 2004). It is understood that the concept of linguistic expression is not uniquely definable since separate elements can have different interpretations. The criterion for the adoption of an interpretation is its correspondence to the real world. A common understanding of the language needs to be attained, such common understanding is a prerequisite for a stable interpretation of the language.
- **Analysis of taxonomical relations.** Concepts unified through taxonomy are analyzed according to their meta-characteristics: identity, rigidity, unity, dependence, thereby revealing more readily the intended meaning of the taxonomical relations. This is the course followed by Christopher Welty and Nicola Guarino (Welty and Guarino, 2001).
- **Methodologies specific to information systems.** Typical concepts of information systems are formalized: the system, the subsystem, unification. Yair Wand and Ron Weber (Wand and Weber, 1990) use the formal model to confirm whether the system is properly divided into components. Mauri Leppänen (Leppänen, 2005) proposes the following methodology for the analysis of the output of information systems – "A system of perspectives is composed of five perspectives. These are the systelogical perspective, the

infological perspective, the conceptual perspective, the datalogical perspective, and the physical perspective." Māris Treimanis (Treimanis, 1998) recommends an aspect-oriented approach when structuring the output of an information system.

Ontologies are developed using various means and differ in the way they depict the world. Standards in ontology are necessary for the regulation of the following:

- What should be included in an ontology
- What are the basic categories and entities
- How are depicted the entities taking into account the knowledge level of the prospective user.

A great variety of backgrounds is needed for the development of ontologies. The author's aim is to develop a methodology for the user who is an expert in the problem-domain albeit without any special knowledge in formalized engineering knowledge systems.

3. ONTO6 methodology

The development of the ONTO6 methodology was influenced by a "6W" approach based on six questions, which, it seems, was first mentioned by the Greek rhetorician Hermagoras already in the year 1 B. C. (Robertson, 1946).

The 6W approach can be considered as a means of obtaining essential information by asking the questions - What, Where, When, How, Why, Who.

The author has named his methodology ONTO6, a name that was chosen not only because ontology is used to define a knowledge model, but also because the development of the ontology was influenced by the 6W approach.

The 6W approach has been adapted to the organization of business knowledge (Malone et al., 2003), the depiction of business structures and enterprise architecture (Sowa and Zachman, 1992; Zachman, 2008), journalism, police work (SixWs, 2022), the organization of brain-storming sessions (Mindtools, 2022), the sphere of architectonic design (Lan, 2004), user modeling (Yudelson et al., 2005), the planning of information systems (Treimanis, 1998; Iljins and Treimanis, 2010), but it is not known to be used in the area of informatization.

The ONTO6 methodology makes use of the 6W framework:

What, Where, When, How, Why, Who.

It is aimed at identifying concepts, determining the interaction between objects corresponding to those concepts and determining the functionality of the objects.

The ONTO6 methodology is based on a semi-informal meta-ontology.

The stages of applying the ONTO6 methodology are:

- the development of a meta-ontology instance appropriate for the domain to be informatized
- the development of an initial ontology from the meta-ontology instance; and
- the gradual detailing of domain concepts that appear in the initial ontology – the development of an enriched initial ontology.

The end result is an ontology cluster, comprising a meta-ontology, a meta-ontology instance, an initial ontology, and an enriched initial ontology.

The ontology cluster is examined for its comprehensibility and its suitability for the domain, thus obtaining answers to several questions of competence.

To achieve a sufficiently general methodology, one that can be applied to the conceptualization of diverse domains to be informatized, a base structure has been incorporated into the methodology as well as a process for obtaining a useful model of the conceptualization of a particular domain from the base structure. This base structure in the ONTO6 methodology is a knowledge model that contains the meta-concept – aspect space. Aspect space describes all possible aspects of the domain to be informatized by grouping them into subsets.

For a given aspect set A = {a1, a2,.,a_i,.. a_n}, where i = 1 to n, where n is a natural number and a_i is an aspect of the domain to be informatized, the aspect space (A) is the set of all the subsets of the aspect set A (power-set). Therefore any element of the aspect space is a subset of an aspect set.

The aspect space remains constant for any domain to be informatized, however a suitable aspect space element must be allocated to the domain. From the knowledge model a usable model for the particular domain to be informatized can be derived. ONTO6 knowledge model is built (see Figure1). Figure1 shows a set of concepts of the domain to be informatized, namely, aspect subset, sub-aspects, concept instances. The relations among them are determined by procedures elicitating sub-aspects and concept instances from the textual information on the domain.

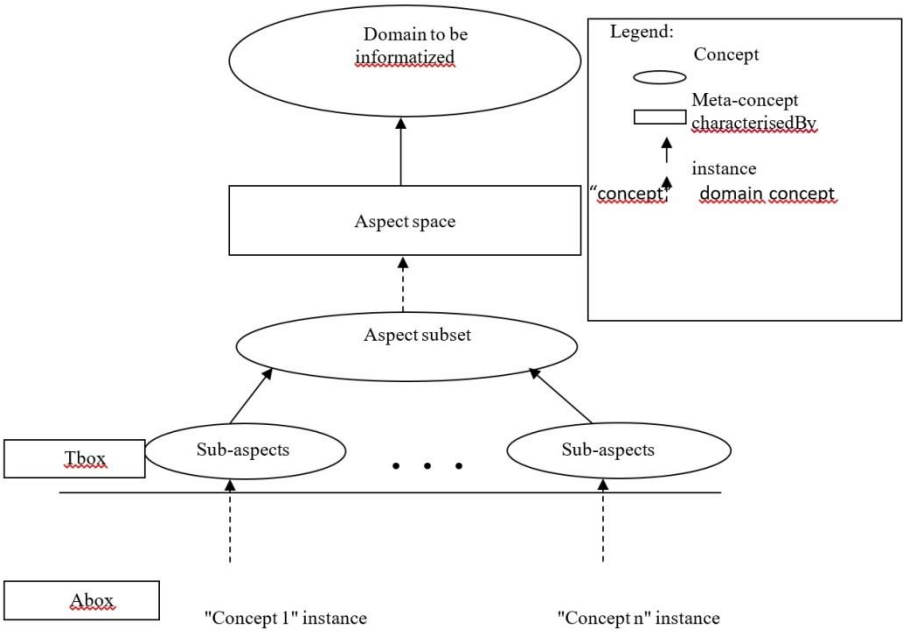

Figure1. ONTO6 knowledge model

In order to obtain the model for the domain to be informatized, a procedure is applied to the knowledge model for determining the aspect subset (an instance of the aspect space) – the frequence of terms corresponding to a particular aspect is calculated, least

frequent aspects are not included in the aspect space. A subjectively chosen threshold value is used for determining the essential aspects; a procedure for adding sub-aspect class instances is developed using text morphological analysis.

In line with the six question approach (Robertson, 1946), the aspect set, A, is chosen to be

A = {What, Where, When, How, Why, Who}.

It is proposed to depict the concepts of the knowledge model in the language OWL with classes. For example, the term “Who” is shown as follows in the syntax of OWL RDF/XML:

```
<owl:Class rdf:about="#Who">
<rdfs:subClassOf rdf:resource="#Aspect set"/>
<rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
>Who</rdfs:label>
</owl:Class>
```

The meta-concept "Aspect space" is depicted as a class of classes with restrictions on the class elements. In OWL RDF/XML syntax this appears as:

```
<owl:Class rdf:ID="Aspect space">
  <rdfs:comment>
    This is the power-set.
  </rdfs:comment>
  <owl:sameClassAs>
    <owl:Restriction>
      <owl:onProperty rdf:resource="&rdfs;#subClassOf"/>
      <owl:hasValue rdf:resource="#Aspect set"/>
    </owl:Restriction>
  </owl:sameClassAs>
</owl:Class>
```

The relationship is shown as a property of the object or data type.

For example, the relationship "characterisedBy" is shown in the syntax of the language OWL RDF/XML as a property of the object "Infodomain" as follows:

```
<owl:ObjectProperty rdf:ID=" characterisedBy ">
    <rdfs:range rdf:resource="#Aspect space"/>
    <rdfs:domain rdf:resource="#Infodomain"/>
  </owl:ObjectProperty>
```

The relationship between the concepts of the knowledge model is depicted in an ontology, which is referred to as a meta-ontology because it contains the meta-concept “Aspect space”, whose instance is the concept “Aspect subset”. Meta-ontology is an essential tool of the ONTO6 methodology. The ONTO6 methodology prescribes the development of a meta-ontology instance in conformance with the domain to be informatized, the development of an initial ontology from the meta-ontology instance and the enrichment of the initial ontology in subsequent informatization. The initial ontology does not change during informatization process. A visualization of the ONTO6 meta-ontology can be built (see **Figure 2**). The circles denote the possible aspects, while the arrows show possible relationships between the aspects. In meta-ontology aspect space instances, some of the arrows between the aspects as well as some aspects themselves may be absent along with the arrows.

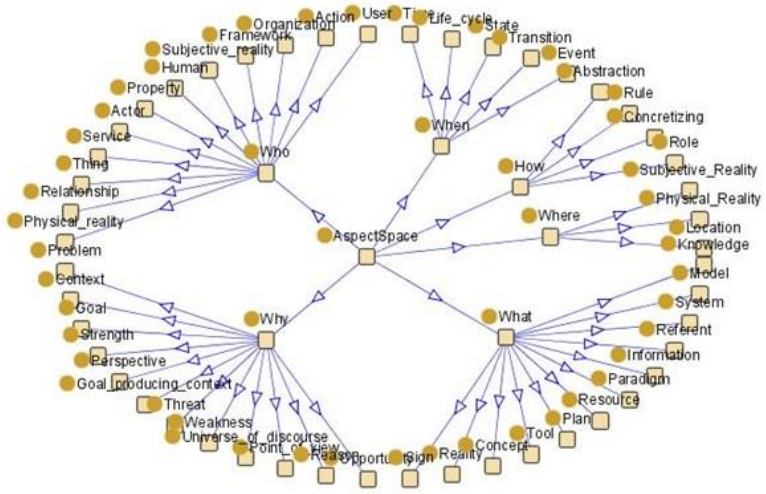

Figure 2. ONTO6 Meta-Ontology Highest level Simplified Visualization

Author's ONTO6 methodology was successfully applied to several domains to be informatized including the Latvian Education Informatization System (LIIS) (Bicevskis et al, 2004).

The ontologies gained as a result of the ONTO6 methodology stages provide answers to questions of competency formulated by necessity in the development of the methodology:

- what are the essential concepts in the given problem domain? (the meta-ontology instance contains only the essential aspects – Who, Where)
- what are the relevant sub-concepts of the essential concepts? (the meta-ontology instance contains some sub-aspects of the essential aspects)
- which aspects of informatization must be examined in more detail? (the initial ontology includes the sub-aspect instances – Abox elements – schools, school boards, ministries, society, educational content, training, administration, infrastructure, information services)
- what kind of functionality is inherent (desired) in the specific aspect? (refined ontologies and visualizations agreed with the user describe in detail the desired functionality)
- which problem domains are similar to the given domain? (it is natural to consider as similar those domains which have the same essential aspects as the LIIS domain).

With ONTO6 methodology it is possible to find essential concepts for different domains.

4. From methodology to framework

Some constraints of ONTO6 methodology usage are: a fixed algorithm for concept elicitation, tiresome manual work to add ontology class instances (individuals) into an ontology, manual comparison of results with expert results.

As an input for domain description mainly textual information is used. The author for the proof-of-concept has used the document (Balodis et al., 1998) containing description of realization of the concept of universal information service in the country, providing every member of the society with quality access to all types of information in accordance with the procedures set forth in the regulatory acts.

The author has looked at several tools which could help at concept elicitation, namely, tools for finding word patterns – AntCone, WordSmith Tools, #LanesBox, SCP, corpkit, TextStat and LVTagger. Author has decided to use the Latvian language text analysis tool LVTagger developed by Peteris Paikens (Paikens, 2022) because the fine-tuning of the essential concepts elicitation according to the particular domain can be easy accomplished using LVTagger.

The author has decided to use the Cellfie Plugin for Protégé 5 (Cellfie, 2022) for automatically entering class instances into an ontology for a particular domain.

As the result the process of adding ontology class instances (individuals) into an ontology was created.

But for the task to learn significant concepts (terms) the input of individuals (class instances) in ontology is only the first step. Important is to calculate the frequence (occurences) of each individual, because the frequencies are used to define the significant concepts.

4.1. Concept elicitation with NLP-PIPE

The author decided to incorporate into the ONTO6 framework's workflow the Latvian NLP Pipeline as a service functionality (Znotiņš and Cīrule, 2018). NLP-PIPE provides a simple and scalable NLP pipeline publicly available under GNU General Public License v3.0 license. NLP-PIPE includes state-of-the-art NLP components for Latvian. NLP-PIPE currently provides following automatic annotations for Latvian: tokenization, morphological analysis, dependency parsing, named entity recognition and coreference resolution (see **Figure 3**). Components are based on already available tools, some with improved models and trained on new larger datasets. A statistical morphological tagger achieves 97.9% accuracy for POS recognition and 93.6% for full morphological analysis, implemented in Java (Grūzītis and Znotiņš, 2018).

Figure 3. Latvian NLP tool pipeline (Znotiņš and Cīrule, 2018)

For a proof-of-concept the author used the public demo site of NLP-PIPE (WEB, a). From the available pipeline tokenization, morphological analysis and dependency

parsing were used for a text of several sentences. The format of the result JSON was chosen. According to ONTO6 methodology the concepts have to be assigned to corresponding aspect from the aspect set {What, Where, When, How, Why, Who}. In Latvian seven declentations (cases) help to choose the appropriate aspect – Nominative, Genitive, Dative, Accusative, Ablative, Locative, Vocative. Not matching parts of speech are ignored.

In **Figure 4** a part of JSON file is shown containing the result of NLP-PIPE morphological analysis of the word "skolās" (Locative case corresponds to the aspect "Where")


```
{
  "deprel": "obl",
  "features": "Skaitlis=Daudzskaitlis|Šķirkļa_ID=298330|Vārds=skolās|Šķirkļa_cilvēklasāmais_ID=skola:1|
  "form": "skolās",
  "index": 12,
  "lemma": "skola",
  "parent": 10,
  "pos": "ncfpl_",
  "tag": "ncfpl4",
  "ufeats": "Case=Loc|Gender=Fem|Number=Plur",
  "upos": "NOUN"
},
```


Figure 4. JSON description of a word in case Locative

To calculate the frequence (occurences) of each individual (in current implementation a single word) the result of morphological analysis is processed with a Python program written by the author generating pairs (word, frequence). The source code is given in **Figure 5**.

```
import json
def findTerms(asp, fn):
  """ The findTerms(asp) function finds terms,
  corresponding to aspect asp in the json file fn,
  returns a Counter object with terms and their frequencies """
  f = open(fn, "r")
  data = json.load(f)
  case = 'Case='+asp
  occur = {}
  for i in data['sentences']:
    for j in i['tokens']:
      if j['ufeats'].find(case)==0 and j['upos']=='NOUN':
        occur[j['lemma']] = occur.get(j['lemma'], 0) + 1
  f.close()
  return (occur)
```

Figure 5. Term frequences calculation

The output of Python program – a Counter object has to be used as an input for process adding ontology class instances to an ontology.

4.2. Adding individuals and frequencies to an ontology

Adding individuals to an ontology can be achieved by Cellfie (needs as an input an Excel spreadsheet obtained from Python Counter object) or with Protégé API.

The Excel spreadsheet in csv form can be generated from Python Counter object by a Python program written by the author (see **Figure 6**).

```
import csv
def createCsv(asp, terms, fn):
  """ The createCsv(asp, terms, fn) function creates csv file fn
  from Counter object terms containing terms for aspect asp """
  f = open(fn, "w", encoding='UTF8', newline='')
  header = ['Individual', 'Class', 'Aspect', 'Occurrences']
  writer = csv.writer(f)
  writer.writerow(header)
  for i in terms:
    data = []
    data.append(i)
    data.append('Location')
    data.append(asp)
    data.append(terms[i])
    writer.writerow(data)
  f.close()

fn1 = "Teik01_ministrija-informatizacija-sodien.json"
terms = findTerms('Acc', fn1) # What
createCsv('What', terms, 'What.csv')
terms = findTerms('Nom', fn1) # Who
createCsv('Who', terms, 'Who.csv')
terms = findTerms('Loc', fn1) # Where
createCsv('Where', terms, 'Where.csv')
```

Figure 6. Csv file generation from Python Counter object

The program takes as an input a json file, generated by NLP-PIPE from an excerpt from National program Informatics document (Balodis et al., 1998). The program calls term frequencies calculation program findTerms for aspects What, Who and Where and generates corresponding csv files. The resulting file Where.csv for the aspect Where is shown in **Figure 7**.

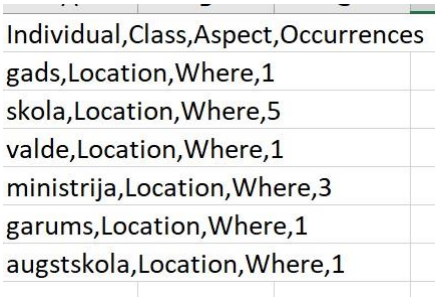

Figure 7. Csv file for the aspect Where

The term frequencies calculation should be carried out for every aspect {What, Where, When, How, Why, Who}.

Conclusion

The ONTO6 methodology has proven to be useful in situations where a compact view is desired of a complicated domain. It has shown itself to be well-suited to the development of a unified user understanding of the domain and for the creation of a description of the essential domain characteristics.

The usage of NLP-PIPE automatic annotations for Latvian texts has made simple the complex task of finding concepts corresponding to aspects. The possibility to analyse non-Latvian texts has to be studied.

The ONTO6 framework will serve as a convenient way to apply the ONTO6 methodology. Further work should be done to fully automate the workflow.

References

Balodis, R., Bārzdiņš, J., Bičevskis, J. et. al. (1998). *National program Informatics* (Latvian). Riga.

Bicevskis, J., Andzans, A., Ikaunieks, E. et al. (2004). Latvian Education Informatization System LIIS, *Educational Media International* 41(1), 43–50.

Cellfie (2022). *Cellfie Plugin*, available: https://github.com/protegeproject/cellfie-plugin/wiki.

Grūzītis, N., Znotiņš, A. (2018). Multilayer Corpus and Toolchain for Full-Stack NLU in Latvian. *CLARIN Annual Conference proceedings*, (8-10 Oct. 2018, Pisa, Italy), pp. 61–65.

Guarino, N. (1998). Formal Ontology and Information Systems, in Guarino, N. (ed). *Formal Ontology and Information Systems.* Proceedings of FOIS'98, IOS Press, Amsterdam, Trento, Italy, pp. 3–15.

Iljins, J., Treimanis, M. (2010). *From Organization Business Model to Information System: One approach and Lessons Learned*, 19th International Conference on Information Systems, Prague, Czech Republic.

Lan, J. (2004). A Preliminary Study of Knowledge Management in Collaborative Architectural Design, *CAADRIA2004*, Seoul, Korea, pp. 35–47

Leppänen, M. (2005). An Ontological Framework and a Methodical Skeleton for Method Engineering. Helsinki.

Malone, T. W., Crowston, K. et al. (eds) (2003). *Organizing Business Knowledge: The MIT Process Handbook*. MIT Press.

Mindtools (2022). *Mindtools. Starbusting template*, available: http://www.mindtools.com/pages/article/newCT_91.htm.

Paikens, P. (2022). *Latvian morphological tagge*r, available at https://github.com/PeterisP/LVTagger.

Robertson, D. W. Jr. (1946). A Note on the Classical Origin of 'Circumstances' in the Medieval Confessional. *Studies in Philology* 43(1), 6–14.

SixWs (2022). *SixWs. Online Encyclopedia Wikipedia*. available: http://en.wikipedia.org/wiki/Six_Ws.

Sowa, J. F., Zachman J. A. (1992). Extending and formalizing the framework for information systems architecture. *IBM System Journal* 31(3), 590–616.

Straujums, U. (2008). Conceptualising Informatization with the ONTO6 Methodology, *Acta Universitatis Latviensis. Computer Science and Information Technologies*, Vol. 733, University of Latvia, Riga, pp.241–260.

Straujums, U. (2010). *ONTO6 Methodology*, Ph.D.Thesis, University of Latvia, Riga.

Straujums, U. (2022). Towards ONTO6 Framework for Concept Elicitation, *CEUR Workshop Proceed*ings, Volume 3158, Riga, pp. 91–100, available at http://ceur-ws.org/Vol-3158/1613-0073.

Treimanis, M. (1998). ISTechnology – Technology Based Approach to Information system Development, *Proceedings of the Third International Baltic Workshop "Databases and Information Systems"*, vol. 2, Riga, pp. 76–90.

Yudelson, M., Gavrilova, T., Brusilovsky, P. (2005). Towards User Modeling Meta-Ontology, *UM2005, LNAI 353*8, Edinburgh, UK, pp.448–452.

Wand, Y., Weber, R. (1990). An Ontological Model of an Information System. *IEEE Transactions on Software Engineering* **16**(11), 1282–1292.

WEB (a). *NLP-PIPE: Latvian NLP Pipeline as a Service*, available at https://nlp.ailab.lv.

Welty, Ch., Guarino, N. (2001). Supporting ontological analysis of taxonomic relationships. *Data and Knowledge Engineering* **39**(1), 51–74.

Wyssusek, B. (2004). Ontology and Ontologies in Information Systems Analysis and Design: A Critique, in *Proceedings of the Tenth Americas Conference on Information Systems*, pp. 4303–4308.

Zachman, J, A. (2008). *Framework2. The Concise Definition*, available at https://www.zachman.com/16-zachman/the-zachman-framework/35-the-concise-definition

Znotiņš, A., Cīrule, E. (2018). NLP-PIPE: Latvian NLP Tool Pipeline. *Frontiers in Artificial Intelligence and Applications*, Volume 307, IOS Press, pp. 183–189.

Received February 20, 2023, revised March 11, 2023, accepted March 16, 2023