# Data Modeling of Complex Historical Information

### Ontology Engineering based on occupational Articles from Early Modern Encyclopedias and Modern Classification Schemes

Robert Nasarek[1]

**Abstract:** This paper describes a strategy for modeling complex historical data using historical professional articles in order to structure, classify, and map information units. Classification systems like the *Historical International Standard Classification of Occupations (HISCO)* were largely built on normative assumptions about occupations in the 19$^{th}$ and 20$^{th}$ century, which led in part to occupational groups, that are inappropriate for the pre-modern workplace. An ontological attempt based on verifiable sources, assisted by statistics and graph algorithms could offer solutions for the deficiencies of classifying approaches. The paper describes a workflow for implement information from unstructured texts into a labelled property graph and gives modeling guidelines for dealing with the multidimensionality of ontological and process-oriented data.

**Keywords:** Data modeling; Germanet; Graph database; Hisco; Labelled property graph; Maxqda; Neo4j; Occupation; Ontology; OCR; Python; Qualitative data analysis

## 1 Introduction

The *European Network for the Comparative History of Population Geography and Occupational Structure 1500–1900* writes on their homepage:

> „Economic historians are drowning in detailed local studies and buffeted by contradictory and methodologically problematic international comparisons based on incommensurable national studies." [Th17]

The *International Labour Office* (ILO) already recognized the problem of the comparison of occupations across countries back in the 1950s. This resulted in the development of the *International Standard Classification of Occupations* (ISCO) – a classification system for governmental use to facilitate international comparisons of statistical data, give guidance to develop own national classification system, and map occupational titles between different languages. [In58, p. 1] The description and classification of occupations is quite useful

[1] Martin-Luther-Universität Halle-Wittenberg, Institut für Geschichte, Emil-Abderhalden-Str. 26–27, 06108 Halle (Saale), Germany, robert.nasarek@geschichte.uni-halle.de, https://ooo.hypotheses.org/

not only for governance, but for a series of social sciences as well. The investigation of social structures, mobility, demography, labour markets and even power relations between people rely on socio-economic circumstances – which are substantially connected with their occupation. On the other hand, the occupation of a person is one of the most common information in serial historical sources like address books, vital statics or tax registers. This demonstrates the desideratum of a valid classification system, especially for historical scientists, who examine socio-economic features. *ISCO* is not intended to deal with historical information, therefore the *Historical International Standard Classification of Occupations (HISCO)* was built open the *ISCO-Version 1968* for coding occupations of the 19th and 20th century. [In19]

Even if (historical) classification systems like *(H)ISCO*, *the Primary, Secondary, Tertiary System* (PST) or the German *Klassifizierung der Berufe 2010* (Classification of Occupations 2010 – KldB 2010) are well constructed for several scientific purposes, they are less suitable for the diversified professional world before 1800. The main issues can be illustrated by assigning the common profession „blacksmith“ to *HISCO*. A *HISCO* entry consists of a six-digit number, a nomenclature and the description of the *HISCO* group. However, *HISCO* does not code individual occupational titles, but the smallest resolution is the so-called *mirco group* (see table 1).

| Number | Name | Description (Task and Duties) Micro Groups |
|---|---|---|
| 83110 | **Blacksmith, General** | Forges and repairs articles of iron and steel, such as hand tools, hooks, chains, agricultural implements and metal structural parts, using hand or power hammers. |

Tab. 1: *HISCO* entry Blacksmith, General

First of all, classification systems like *HISCO* did not use verifiable empirical data to build occupational groups. *HISCO*, like most other classification schemes, was formed on normative assumptions about occupational titles and therefore has no verifiable structured or statistical database.[2] ([vLMM02, p. 9–15 and 25–30][Na16, 1–19]) In some cases this leads to incorrect or incomplete results. Comparing the description of the blacksmith in HISCO with the corresponding lemma „Schmid, Schmidt, Schmied“ within Johann Heinrich Zedlers *Grosses vollständiges Universal-Lexicon aller Wissenschaften und Künste* (Great Complete Encyclopedia of All Sciences and Arts) it lacks a significant activity:

> „[... a blacksmith] also must have a thorough understanding of equine medicine, be able to judge all deficiencies and accidents of horses, and treat

[2] The German system *Classification of Occupations 2010* (Klassifizierung der Berufe 2010) pursues a more empirically oriented approach by assigning competences to occupations according to which the groups are clustered; see also [Bu11, p. 33–38]. Even if the approach is a step in the right direction, there is no access to the (sources of) competence profiles. The sheer mass of over 7000 competences (which additionally are not in a taxonomical order) also makes it difficult to classify them evenly.

> them with medication, externally by bleeding and laying on good horn healing and other ointments and plasters, or internally with powders and enemas.“ [Ze54, vol. 35, p. 361]

In addition to that, the categories (tasks, skills, products, materials and so on) and their values (forges, chain, iron, hammer etc.) which led to the grouping of occupational titles are not tagged in the profiles. This impedes specialised queries and leads to an inflexible system. In this way, it is difficult to query something like: „Who works with metal?“ or „Who uses hammers?“ Furthermore the word-processed mapping of occupations – that means: finding a suitable occupational class within over 1600 possible micro groups, reading the description and decide if it matches – lowers the intercoder-reliability and takes a lot of time. To remedy such problems, the database used in the following project bears four features:

1. The data of the occupational profiles relates to reliable resources – historical encyclopedias include all necessary information and serve quite a lot of occupational categories to build dense profiles.
2. The information is modeled as an ontology[3] and is implemented into a labelled property graph[4] – ontologies allow a much more sophisticated approach to complex semantic data then word-processed formats.[5]
3. The occupational groups are build on statistics and graph algorithms – since the data is well organised, statistical computing can render more valid results.
4. The data should be accessible over a web interface – *Representational state transfer* (REST) as well as database dumps (*.sql, *.csv) should be available for users.

## 2 From unstructered text to a labelled property graph

The primary objective is to develop an ontology for approximately 200 historical occupational titles.[6] The work will not be able to provide an exhaustive collection of all existing premodern occupations titles. Nor can all aspects of an occupation be considered equally (issues such as law, traditions or theology will not be implemented). The project will only focus on the most commonly used categories in the sources and in the modern classification schemas: productive activities, education and qualification, aptitude, and working conditions. These categories are used to create occupational profiles from the sources and *HISCO*

[3] For a detailed insight into the topic of ontology engineering, see [Gr93, GOS09, NM01].
[4] For a short introduction see [RWE15, p. 25–64]
[5] Attemps like hisco2rdf transpose *HISCO* to the *Resource Description Framework* (RDF) and *Simple Knowledge Organization System* (SKOS), but don't increase the level of formalisation within the descriptions required for automatic mapping or statistical clustering.
[6] A occupational **title** is only the appellation of an occupation; an occupational **profile** contains structured information about activities, materials, products, etc. In this way, a occupational title can have different profiles depending on the source.

descriptions. Within this framework, three contributions can be made: By comparing the profiles from the sources with the *HISCO* profiles, the validity of *HISCO* in relation to pre-modern professions is checked. Furthermore, it can be the strategic starting point for a methodically clean and cheaply scalable ontology of historical occupations. In general, it produces insights for data modeling of unstructured but highly complex information, as often found in digital humanities projects.

### 2.1 Sources and Data Foundation

The main historical sources are articles from Johann Heinrich Zedlers *Grosses vollständiges Universal-Lexicon* [Ze54]. It contains 63.000 two-columned folios with ca. 284.000 articles, which were compiled by various (mostly unknown) authors. Even if those compilations mainly contains identical quotations from other sources, without reflections or syntheses by the compiling authors; the editors of those articles condensed the available knowledge, redacted the material, removed redundant information and arranged the content in a meaningful manner. [Sc13, p. 41] Along with their corresponding description counterparts in *HISCO*, this will result in ca. 400 occupational profiles.
The first question was which occupations should be used to build the ontology. Since the occupations had to be more or less typical for the 18th century, it was reasonable to use professional titles out of contemporary books of estates. For this reason, the selection of the occupation titles relies on Christoph Weigels *Abbildung Der Gemein-Nützlichen Haupt-Stände* (Illustration of the useful main estates) [We98]. The book contains 221 „professions“, from „Advocat“ (advocate) to „Zuckerbacher“ (confectioner). Since Weigels descriptions are much longer compared to the articles in Zedlers *Encyclopedia*, but contain less useful information in the sense of the project, Zedler was chosen as the main source for building the occupational profiles. [Na16, p. 60–61] Not every „profession“ in Weigels *Illustrations* could be found in Zedlers *Encyclopedia*, the number of occupational titles to be examined is currently 178. Zedlers *Encyclopedia* is online available only as a digital copy hosted by the *Munich DigitiZation Center*; but in order to perform further annotation and transformation, they must be available in full text. The Optical Character Recognition (OCR) engine *Ocropus* was used for this purpose.[7]

### 2.2 Qualitative Data Analysis

In this project, as is often the case in digital history, the data model emerges while working with the sources. Due to the complexity of the texts (different writers deal with a diverse but changing spectrum of topics with varying information depth) it is more elaborate to develop a solid data model. The risk of „proliferation“of the model or overlapping meanings of entities is higher. Especially when dealing with unstructured but informative texts, it is

[7] In the meantime *Ocropus* is embedded in the quite more sophisticated OCR-framework *OCR4all*.

important to develop entity definitions that are exhaustive, clearly distinguishable, consistent and easy to communicate. ([Ku16, p. 93][SW09, p 80–82 and 146–147]) The method of qualitative structured content analysis developed by [Ku16] offers an excellent approach to gain domain knowledge about the occupations. Content analysis is used to categorise the text segments of the articles and structure their information.[8] Since the main concepts for classifying occupations in *HISCO* are tasks and duties [In69, p. 5][9], the most important code is called „productive activity“and is applied to segments that deal with the manufactured products, used materials and tools or the already mentioned tasks. Other a priori concepts were education and qualification, aptitude, taxonomy, working conditions and occupational status. Beside those main classification concepts, the articles from ZEDLERS *Encyclopedia* show a number of other ideas, with which a occupation is described, but which are not implemented in the database. The most common are „narratives“of theological derivation or advices to recognize good quality or fraud, „regulations“in the form of conventions and laws concerning the exercise of the profession or „community measures“in the sense of information about traditions and the like (see table 2).

| Code | Occurrence |
|---|---|
| Productive Activity | 96% |
| Taxonomy | 89% |
| Working Conditions | 48% |
| Education | 42% |
| Aptitude | 12% |
| Narrative | 48% |
| Community Measure | 27% |
| Regulation | 27% |
| Prestige | 18% |
| Occupational Status | 11% |
| Others | 68% |

Tab. 2: Occurrence of concepts in the 178 descriptions of historical occupations. Multiple occurrences of codes are counted only once per article.

The coding process was carried out in two phases. In the first phase, the coding unit was set to the length of a sentence and the deductively developed main codes were applied to the articles (productive activity etc.); but also new codes were developed inductively from the material (regulation, community measure etc.). The software *MAXQDA* was used for this stage. *MAXQDA* is a very productive and user friendly tool to handle a great amount of unstructured text at the stage of *building* the domain knowledge. Besides annotation, *MAXQDA* provides features to assemble codings and display them in one single prompt (see figure 1).[10]

[8] For a quick, but more detailed introduction into qualitative data analysis, see [Mü19].
[9] *HISCO* inherits it’s systematic from *ISCO 1958* [In58] and *ISCO 1968* [In69].
[10] An alternative approach could be a markup with *Extensible Markup Language* (XML), but the effort to transform, transpose or assemble tags and contents was disproportionately high in comparison with *MAXQDA*.

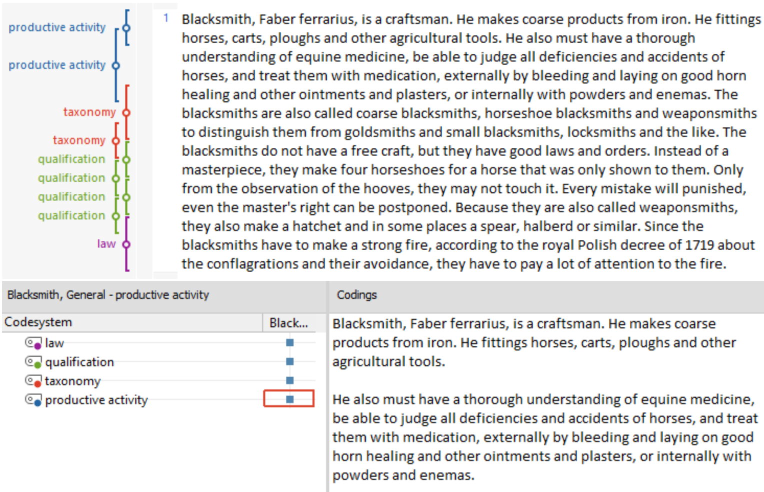


Fig. 1: Annotation and reassembling codes in *MAXQDA*

In the second phase, the main codes with a length of a sentence are divided in sub codes, as „task“, „product“, „material“ and so on as in the case of the productive activity. The information of the codings (i. e. „forge“, „chain“, „silver“ is normalised into a table structure of an excel sheet. During the process, the data undergoes deeper refinement, standardisation and enrichment with metadata.

Since the collected data should be reusable in Katrin Moellers „Ontologie historischer deutschsprachiger Berufs- und Amtsbezeichnungen“, the tabular data model represents a structured form of the syntax of the coding. This should guarantee a „human readable“ form as well.[11] It is not necessary to normalise the data up to third normal form, because the table will only be used as an import schema for the labelled property graph database management system *Neo4j*. The tabular data model will satisfy functional requirements if every value is standardised and every tupel is in a parseable shape.

### 2.3 GermaNet and Neo4j

The next phase of the project deals with the ambiguous side of language and the semantic relationships between lexical units. For example the word „bat“ could be an an implement used to hit a ball or could be a nocturnal flying mammal. On the other hand „candy“ has the same meaning as „sweet“. To disambiguate those cases, every lexical unit needs a link to a semantic unit. The German wordnet *GermaNet*[12] provides a solution for this

[11] Current schema can be found here.
[12] Oriented on the English predecessor *WordNet*.

kind of problem: „GermaNet is a lexical-semantic net that relates German nouns, verbs, and adjectives semantically by grouping lexical units that express the same concept into synsets and by defining semantic relations between these synsets. “ [Se18]. That means, that *GermaNet* not only provides semantic identification for words, but delivers a relational network of hyponyms, hypernyms and synonyms. This offers the opportunity to easily add a taxonomic dimension to enable more sophisticated queries.

The data from *GermaNet* is stored in the database management system *PostgreSQL*. With a script using the *Python* libraries *SQLAlchemy* and *Pandas*, it is possible to enrich the structured data with lexical information from *GermaNet*.
Unfortunately, the sources contain a variety of specific words that are not available in *GermaNet*. So in a first run, every word that was present in the initial database was identified through it's lexical unit id in *GermaNet*, during which unknown words were collected and then added to the local instance of *GermaNet*; after that another run was performed to add the new lexical unit ids to the specific words in the excel file.

The ontology is implemented in the database management system *Neo4j*. As the self-titled „World's leading graph database“ [Ne15] *Neo4j* is a full featured infrastructure administer a labelled property graph, it supports many programming languages, most importantly *Python*, which is used at several positions in the project. In addition to that, *Neo4j* has a vibrate and comprehensive learning section, together with several well documented libraries like *Graph Algorithms* or *APOC*.
*Neo4j* provides a *Graph Algorithms* library, which is used to compute metrics for graphs, nodes or relationships. Predefined algorithms for community detection, path finding or similarity could help to cluster occupational groups. Furthermore, the *APOC* library consists of about 450 (more specialised) procedures and functions to support different tasks. This brings the advantage to build upon reliable structures and avoid costly reinvention.

To determine the similarity of occupations, their characteristics are compared within the main codings (i. e. productive activity: „Forge silver into chain“ or „Mould gold into ring“). The shortest path between two subcodings (i. e. „forge“ and „mould“, „silver“ and „gold“ and so on) is determined and the common node with the lowest depth in the net is divided by the maximum depth of the net. The mean is formed from the values so that the similarity of the two considered main categories is obtained. These values are also summed up and the average is calculated. This process is carried out with all occupations, so that each occupation has a similarity value to each other. The result is used to cluster the occupations.

Another reason to use a property graph like *Neo4j* and not a semantic graph implemented in the *RDF* and it's extensions, is the possibility to assign literals to properties. [Ba16a] This offers the opportunity to connect meta data directly to the properties and avoids an excessive build of statements over statements, called „reification “ [HM17]; the *SPARQL Protocol And RDF Query Language* (SPARQL) and most triple stores does not yet intrinsically support this feature. [Ca19]

### 2.4 Three Dimensions of Modeling

In order to compare the job descriptions, three dimensions are taken into account in the data modeling. The first dimension concerns the semantic components in compound words. In the German language, word composites are a common form of word formation; several meaningful lexemes or morphemes are combined into one word, such as „Gold“ and „Ring“ into „Goldring“. The first element „Gold“ defines the second element „Ring“ more closely, so semantically the „Goldring“ forms a subclass of „Ring“. [Ba16b, p. 675–676] The already mentioned second dimension of taxonomy concerns the relation of the word classes to each other. If only compounds were considered taxonomically, the part of the first compound member (gold is precious metal is metal etc.) would not be taken into account. Therefore, compounds have to be decomposed into their components, if they contain lexemes relevant for the comparison. This is the case with materials such as the „Gold“ in „Goldring“, but also with formats or topics such as „Physik“ in „Physikexperiment“. In the **lexical-semantic** dimension, the meanings of the words are segmented on the horizontal level, while in the **taxonomic** dimension the respective segments are semantically related on the vertical level.

In addition, the same words in the **analytical** dimension can be related to different codes. For example, a blacksmith uses a hammer to forge chains, while the toolmaker makes hammers. In the case of the blacksmith, the hammer is assigned with the code „tool“, while in the case of the toolmaker the hammer is encoded with „product“. Depending on the context, a word can therefore have different roles, which are an expression of a **relationship** between entities rather than a property or classification of an entity itself. This can result in ambiguities if the dimensions have the same values, but are not modeled separated from each other. Thus, the hammer can belong to the class „tool“ in the taxonomic dimension, but have assigned the code „tool“ on the analytical dimension as well. It is therefore convenient to adapt the dimensions to the different resources of the used meta model. *Neo4j* provides nodes, labels, relationships and properties as slots for the different dimensions. [Ne19] Static and individual information like lexical unit, orthographic form and word definition were placed on the properties of nodes; labels are used for class assignments, like synsets or main codes, and contextual information such as the subcodes („roles“) of words within a statement (hyponym, hypernym, task, product etc.) were implemented as relationships (is_hyponym, is_hypernym, has_task, has_product, etc.). In this way, the information is organized into a formal, logical and structured form of an ontology implemented in a labelled property graph (see figure 2).

### 2.5 Project Status

A local database offers limited accessibility to an interested community, if it is not embedded in a broader framework. The web application framework *Flask* allows to build a website, which provides basic features like querying a database and display the results. Together with

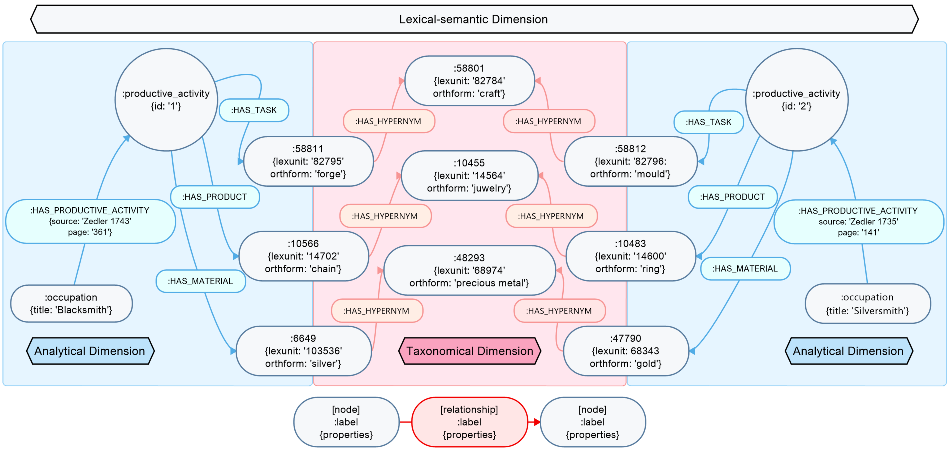


Fig. 2: Three dimensions of information implemented through the different resources of *Neo4js* meta model.

*D3.js*, a *JavaScript* library for manipulating documents based on their data, it is possible to visualise the results and enable access to an audience, who has less experience in *Neo4j*. The data should be accessible over a *REST Application programming Interface* as well as database dumps in common formats like .sql or .csv.

Currently, the project is at a point where all raw data has been transformed into a tabular form and all codings have been linked to the lexical units of the *GermaNet* (if possible).[13] The routines of the graph algorithms were developed and tested on the existing data and provide plausible, but partly too fuzzy results. This is especially the case when occupations have only few data, their characteristics are only roughly described or show a wide spectrum. This problem could be solved if more encyclopaedias were consulted that might contain further or more precise information. What have to be done now is to add missing words and their semantic relations to the *GermaNet* to gain better results and develop a user interface to browse and access the data.

## 3 Conclusion

Occupational classification systems like *HISCO* are largely based on normative assumptions about occupational groups and do not have an empirical database and thus no verifiable structured or statistical basis, which leads to incorrect or incomplete results in some cases – especially for pre-modern occupations. In addition to that, the categories used to group the occupations are not tagged in the profiles. This impedes specialised queries and leads to

[13] Modeling, programming and content-related issues arising from or around the ongoing process are documented on the website ooo.hypotheses.org/.

an inflexible system. An ontology based on verifiable data offers the opportunity to render more valid results.

Although ontology engineering is mainly associated with semantic graphs in *RDF*, ontologies are first and foremost a way to organise and formalise knowledge and are therefore not necessarily to be implemented in *RDF*. The labelled property graph *Neo4j* offers the advantage to connect literals to properties which avoid undue reification and is a good slot to place meta data. Soft factors as the extensive documentation of the features and prefabricated graph algorithms and modules helps to achieve results efficiently and cheaply. This is a not negligible advantage for small projects with few but interdisciplinary working staff, as is often the case in the field of digital humanities.

During the modeling of complex and partly process-oriented data from unstructured texts, it is advantageous to basically consider three dimensions of the data and distribute them among the different resources of the meta model to avoid ambiguity and to simplify queries: the lexical-semantic dimension for individual and context independent information should be realised in properties of nodes; the taxonomical dimension for class assignments should be implemented as labels, and the analytical dimension for context dependent information should be transposed in relationships. Along with entity definitions that are exhaustive, clearly distinguishable, consistent and easy to communicate, the likelihood increases of having developed a solid data model instance that delivers valid results.

## References


[Ba16a] Barrasa, Jesús: RDF Triple Stores vs. Labeled Property Graphs: What's the Difference? `https://neo4j.com/blog/rdf-triple-store-vs-labeled-property-graph-difference/`, 2016. visited on 2019-08-05.

[Ba16b] Barz, Irmhild: Komposition. volume 4. Dudeverlag, Berlin, chapter Das Wort, pp. 135 – 775, 9 edition, 2016.

[Bu11] Bundesagentur für Arbeit: Klassifikation der Berufe 2010: Systematikband, volume 1. Bundesagentur für Arbeit, Nürnberg, 2011.

[Ca19] Cagle, Kurt: Graph Databases Go Mainstream. `https://www.forbes.com/sites/cognitiveworld/2019/07/18/graph-databases-go-mainstream/amp/?__twitter_impression=true`, July 2019. visited on 2019-07-23.

[GOS09] Guarino, Nicola; Oberle, Daniel; Staab, Steffen: What Is an Ontology? In (Staab, Steffen; Studer, Rudi, eds): Handbook on Ontologies, International Handbooks on Information Systems, pp. 1–17. Springer, Dordrecht, 2009.

[Gr93] Gruber, Thomas R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2):199–220, 1993.

[HM17] Hayes, Patrick; McBride, Brian: RDF Semantics, Reification. `https://www.w3.org/TR/rdf11-mt/#reification`, October 2017. visited on 2019-07-25.

[In58] International Labour Office: International Standard Classication of Occupations. Geneva, 1958.

[In69] International Labour Office: International Standard Classication of Occupations. Geneva, 1969.

[In19] International Institute of Social History: HISCO - History of Work. https://socialhistory.org/en/projects/hisco-history-work, 2019. visited on 2019-02-06.

[Ku16] Kuckartz, Udo: Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung. Juventa, Weinheim, Basel, 3 edition, 2016.

[Mü19] Müller, Andreas: What is Qualitative Data Analysis. https://methodos.hypotheses.org/275, 2019. visited on 2019-02-10.

[Na16] Nasarek, Robert: Tätigkeitsprofile in der Frühen Neuzeit. HISCO-Klassifikation vormoderner Berufe anhand qualitativer Datenanalyse. Master's thesis, Martin-Luther-Universität Halle-Wittenberg, Halle (Saale), 2016.

[Ne15] Neo4j: Developer - Neo4j Graph Database Platform. https://neo4j.com/developer/, 2015. visited on 2019-02-10.

[Ne19] Neo4j: Chapter 2. Graph database concepts. https://neo4j.com/docs/getting-started/current/graphdb-concepts/, June 2019. visited on 2019-07-31.

[NM01] Noy, Natalya F.; McGuinness, Deborah L.: Ontology Development 101: A Guide to Creating Your First Ontology: Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.

[RWE15] Robinson, Ian; Webber, Jim; Eifrem, Emil: Graph databases: new opportunities for connected data. O'Reilly, Sebastopol, CA, 2nd ed. edition, 2015.

[Sc13] Schneider, Ulrich Johannes: Die Erfindung des allgemeinen Wissens: Enzyklopädisches Schreiben im Zeitalter der Aufklärung. Akademie Verlag and de Gruyter, Berlin, 2013.

[Se18] Seminar für Sprachwissenschaft Tübingen: GermaNet - An Introduction. http://www.sfs.uni-tuebingen.de/GermaNet/, 2018. visited on 2019-02-10.

[SW09] Simsion, Graeme C.; Witt, Graham C.: Data modeling essentials. Morgan Kaufmann, San Francisco, California, 3 edition, 2009.

[Th17] The Cambridge Group for the History of Population and Social Structure: The European Network for the Comparative History of Population Geography and Occupational Structure 1500-1900 (ENCHPOPGOS), 2017.

[vLMM02] van Leeuwen, Marco H. D.; Maas, Ineke; Miles, Andrew: HISCO: Historical international standard classification of occupations. Leuven University Press, Leuven, Belgium, 2002.

[We98] Weigel, Christoph: Abbildung Der Gemein-Nützlichen Haupt-Stände. Regensburg, 1698.

[Ze54] Zedler, Johann Heinrich, ed. Grosses vollständiges Universal-Lexicon aller Wissenschaften und Künste. Zedler, Johann Heinrich, Halle, Leipzig, 1731–1754.