Tamper, Minna; Leskinen, Petri; Tuominen, Jouni; Hyvönen, Eero

Modeling and publishing Finnish person names as a linked open data ontology

Published in:
CEUR Workshop Proceedings

Published: 01/01/2020

Document Version
Publisher's PDF, also known as Version of record

Modeling and Publishing Finnish Person Names as a Linked Open Data Ontology

Minna Tamper[1,2][0000−0003−1695−5840], Petri Leskinen[1][0000−0003−2327−6942], Jouni Tuominen[1,2][0000−0003−4789−5676], and Eero Hyvönen[1,2][0000−0003−1695−5840]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland and
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
http://seco.cs.aalto.fi, http://heldig.fi, firstname.lastname@aalto.fi

Abstract. This paper presents an ontology and a Linked Open Data service of tens of thousands of Finnish person names, extracted from contemporary and historical name registries. The repository, first of its kind available, is intended for Named Entity Recognition and Linking in automatic annotation and data anonymization tasks, as well as for enriching data in, e.g., genealogical research.[3]

1 Introduction

Actor ontologies of people, groups, and organizations (e.g., Getty ULAN[4]), also called authority files [11] in Library Sciences, are a key ingredient needed in publishing and using Cultural Heritage (CH) Linked Data on the Semantic Web. For representing actor ontologies, there exists several vocabularies, such as FOAF[5], REL[6], BIO[7], and Schema.org [6]. Actor ontologies make a distinction between language-neutral concepts (resources identified by IRIs) and their literal names. In Resource Description Framework (RDF)[8]-based modeling in use on the Semantic Web, only resources can have properties while literal names are considered only atomic data that do not have properties, except a possible datatype and language tag attached. However, in many cases also literal words can have qualifiers and properties: names of things change in time and in context, e.g., female names due to marriage, or the language version form of the name in different countries and cultures (e.g., "Gabriela" vs. "Gabriele"). In linguistic Linked Data repositories [22], modeling phenomena related to the properties of words instead of real world things is actually the main reason for the research. For modeling phenomena like this, the SKOS recommendation has been extended to

[4] http://www.getty.edu/research/tools/vocabularies/ulan/about.html
[5] http://xmlns.com/foaf/spec/
[6] http://vocab.org/relationship/
[7] http://vocab.org/bio/
[8] https://www.w3.org/RDF/

SKOS-XL[9], allowing specifying properties for literal SKOS labels, and various linguistic ontology models such as Lemon[10] and OntoLex-Lemon[11] have been devised for representing linguistic Linked Data repositories.

A person name individualizes and identifies an individual. A person name ontology is a collection of contemporary and historical person names in a machine-understandable way. It is a knowledge graph describing names, their features, and usage in different datasets. In actor ontologies of people, names are often represented as literals. The features of the name are often ignored when describing people in actor ontologies although the name can carry information about its bearer such as socioeconomic status or gender.

This paper introduces a data model for representing person names as an ontology, based on tens of thousands of person names from contemporary Finnish name registries, including also historical names extracted from various CH linked data sources. The new Finnish Linked Open Data name ontology HENKO[12] has been used as a basis for named entity recognition (NER) and linking tasks [7] in automatic content annotation [29] and data anonymization services [25], as well as enriching linked data for applications, such as genealogical network analysis [16,20]. To foster the reuse of the data, this repository of Finnish person name data, first of its kind available, is published as a Linked Open Data service for application developers to use under the open CC BY 4.0 license.

2 Datasets

The data for the person name ontology HENKO was collected from multiple registries. It consists of given and family names and the number of users per name. The amount of users for the given names was calculated by gender. In addition, the given names data included the sum of users who have it as a first and as other given name. The collected datasets, the total number of names, and number of unique names in the data are shown in Table 1.

The first dataset in the table is from the Finnish Digital Agency[13] (FDA), a governmental agency that promotes digitalization of society, secures the availability of data, and provides services for the life events of its customers. The agency publishes Finnish name data as open data in the governmental publication portal avoindata.fi[14]. This dataset contains given names that are used by a minimum of five persons, and family names for the minimum of 20 persons. There are in total 23 018 family names, 9507 male given names, and 12 304 female given names (cf. Table 1). According to the product manager of FDA, the dataset contains only a fraction of Finnish person names. The full registry

[9] `https://www.w3.org/TR/skos-reference/skos-xl.html`
[10] `https://lemon-model.net`
[11] `https://www.w3.org/2019/09/lexicog/`
[12] The name comes from the Finnish name Henkilönimiontologia (Person name ontology); Henko is also a diminutive form of the name Henrik.
[13] `https://dvv.fi/en/individuals`
[14] `https://www.avoindata.fi/data/en_GB/dataset/none`

contains a total of 293 367 family names and 126 119 given names. According to FDA, the names used by less than the given amounts, are not publicly available because rare names can single out individual persons violating their privacy. Most of these unique names come from foreigners, and the rarer Finnish given names are often compound or coined names. FDA publishes the data twice a year; our the data has been collected starting from August 2018.

Dataset	**Family names**		**Given names**			**Total**
	unique	total	unique	female	male	
The Finnish Digital Agency	16 931	23 018	18 206	11 093	8299	42 410
BiographySampo	1205	5535	805	1705	1761	9001
Norssi High School Alumni	1002	4598	233	509	1039	6146
AcademySampo	6721	11 016	946	1389	1423	13 828

Table 1. Amount of names by dataset

In addition to using the FDA data, our ontology has names extracted from the datasets Norssi High School Alumni on the Semantic Web [9], BiographySampo [10], and AcademySampo [17][15]. AcademySampo contains names of university students from 1640 to 1899, and it contains plenty of historical, often Latin-based, names. BiographySampo data is based on 13 100 biographies of significant Finns throughout the history from the 3rd century to present time, and it has many Swedish names used by nobility and upper class because until 1809 Finland was an integral part of Sweden. The Norssi Alumni dataset records students in a Finnish school from 1867 to 1992 and the unique names in it are mostly rare Finnish names. Altogether these datasets provided 15 975 distinct family names, 2791 male and 2500 female names.

In order to have more features for the names in the ontology, the name datasets were processed and enriched using natural language processing (NLP) methods. Family names, for example, can contain nobiliary particles or suffixes. In Finnish family names [26] nobiliary particles are not used, but the names have suffixes that have indicated once if a person came from a place (e.g., suffixes *-la*, *-lä*), or person's socioeconomic status (e.g., scholars, soldiers, clergy with suffixes *-er*, *-ius*). To make this information explicit, the particles and suffixes were extracted from the names. For the particle extraction, the corpus of particles (in other languages) was compiled from the website of the Institute for the Languages of Finland (Kotus)[16] to identify names that contain particles. The extraction of suffixes was done using Lexical Analysis Service's (LAS)[17] [18,19] language recognition service, hyphenation service, and a manually compiled stopword list of words in Finnish and Swedish compound names

[15] `https://seco.cs.aalto.fi/projects/yo-matrikkelit/en/`
[16] `http://www.kielitoimistonohjepankki.fi/ohje/65`
[17] `http://demo.seco.tkk.fi/las/`

(e.g., fi. *Mansikkamaa* eng. *strawberry field*). The process first filters out names ending with a stopword, then detects the language, and lastly hyphenates the name. The last syllable is recorded as the suffix. Short names with only two syllables were ignored because they rarely end with a suffix.

In addition the NLP methods were used in identifying patronymics (e.g., *Jaakonpoika*, eng. *son of Jaakko*) and matronymics (e.g., *Liisantytär*, eng. *daughter of Liisa*). The matronymics and patronymics are identified in Finnish, Swedish, and Russian. In Finnish and Swedish they end with a word that indicates if its owner is a female (sv. *-dotter*, fi. *-tytär*) or a male (sv. *-son*, fi. *-poika*) whereas the Russian counterparts have a gendered suffix (e.g., *-ov*, *-ova*). The preceding part of the word is a person name typically in the genitive case and it can belong to an ancestor of the person. The ancestor's name wass extracted from the preceding part and baseformed with the LAS lemmatization tool. Afterwards, the ancestor's name is used to find names with the same string form. If the name exists, the application identifies the gender by using the existing data. The name instance is typed as matronymic or patronymic depending on the result.

Lastly, the names of HENKO data were linked to their counterparts in DBpedia and Wikidata to enrich the data with etymological information and relations to other names. The names were also linked to the bearers of the names in the source datasets (AcademySampo, BiographySampo, and Norssi Alumni). In addition, the family names can reference place and vocation names [26]. To identify names that refer to places and vocations, the names were linked to the YSO places ontology[18] (Finnish and Swedish place names) and to the Finnish historical occupations ontology AMMO [15]. This information is not only interesting topical information but can be used in tasks such as linking based NER to identify names that can be place or vocation names.

3 A Data Model for Person Names

The data model for person names in HENKO has been created based on the enriched name data. The data model is depicted in Fig. 1. The model has a class for the written representations of the name, the *WrittenNameForm*, that includes the string presentation of the name. Its instances are also instances of the CIDOC CRM's class *E41_Appellation* in order to enable the modeling of names and their alternative forms. This is needed, for example, if a name is translated from Russian to Finnish, as was the case with the Russian tsars *Alexandr I-III*, that were called in Finnish *Aleksanteri I-III*. The *WrittenNameForm* class connects to the *GivenName* and *FamilyName* classes via *isGivenName* and *isFamilyName* properties accordingly.

The *GivenName* and *FamilyName* classes are subclasses of the *Name* class. The *Name* class describes the basic features of the names, such as properties for linking both names to their equivalent representations in other ontologies, to people in other actor ontologies with the same names, in case of compound

[18] `https://finto.fi/yso-paikat/en/`

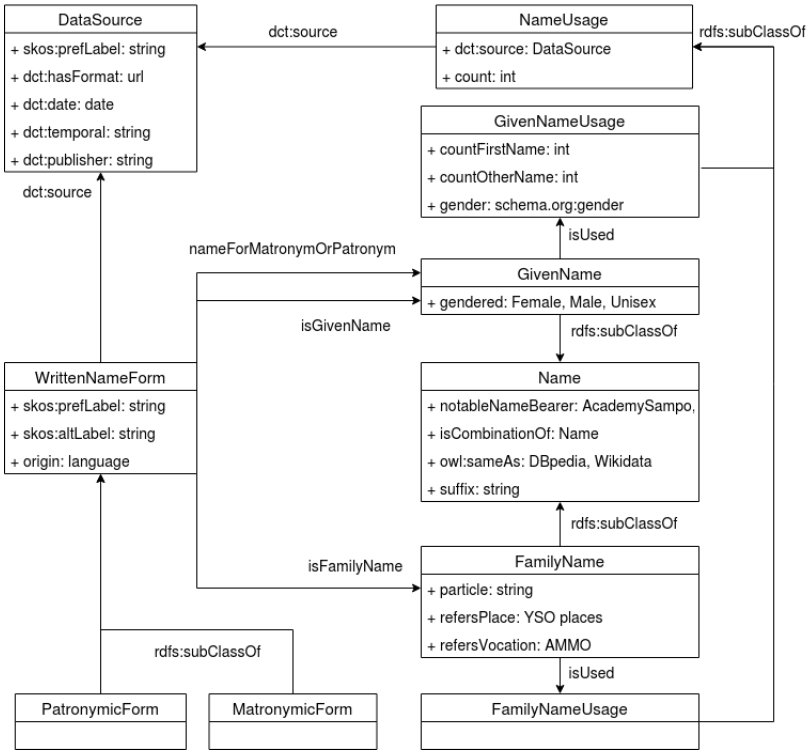

Fig. 1. The datamodel for Finnish person name ontology HENKO.

names by linking it to the parts (e.g., the name *Henna-Maria* can be linked to *Henna* and *Maria*), and to linguistic information, such as name suffixes. Like in the Wikidata [4] model, the *GivenName* class has information about the gender (male, female, unisex) that is inferred for each instance based on the name usage data. The *FamilyName* class instances contain information about the nobiliary particle, such as *von*, or *de la*. Initially, the OntoLex-Lemon [21] and MMoOn [14,13] ontologies were considered for modeling the particles and affixes, but the models did not fit the needs of HENKO because of being too complex or lacking in features to represent them. In addition, the references to places and vocations have been recorded using their own properties. Middle names[19] are not common in Finland and are ignored currently in the processing.

The *GivenName* and *FamilyName* classes are connected to the *GivenNameUsage* and the *FamilyNameUsage* classes through the *isUsed* property. These classes describe the calculated usage of the name. They are the subclasses of *NameUsage* class. The *NameUsage* class describes the general characteristics of its subclasses, such as count (how often a name is used) and source (data source for the information). The *GivenNameUsage* class also separates whether the name has been used as a first name or other name (second, third) in ad-

[19] `https://en.wikipedia.org/wiki/Middle_name`

dition to having the gender attribute. The *DataSource* class, that connects to the *NameUsage* superclass, describes the used sources in more detail. It includes attributes such as date (creation time of the data), URL (where the data was retrieved), temporal information about the dataset, its publisher, and name. The *DataSource* class is also connected directly to the *WrittenNameForm* class.

Finally, the *MatronymicForm* and *PatronymicForm* classes are subclasses of the class *WrittenNameForm*. If the instances of the *WrittenNameForm* class have been identified as patronymics or matronymics, the *WrittenNameForm* instances are complemented with information about the origin (Finnish, Swedish, Russian) and are linked to the given name of the ancestor (*GivenName* class instance) using using Wikidata's property "patronym or matronym for this name". The suffixes from the Russian origin names are recorded using the *Name* class property *suffix*.

4 Use Cases

This section presents the applications of HENKO in automatic annotation tasks. The applications are available as part of the SeCo Text Annotation Service[20].

Gender Identification Service The code behind the service[21] has been developed in the projects Norssi Alumni, BiographySampo, and AcademySampo to determine the gender by person's name. The service uses HENKO vocabularies of given names containing the frequencies of how often each name appears as a male or a female name.

The decision is based on the standard Bayesian approach described in equations 1 and 2. Equation 1 defines the probability $\rho(\gamma|n)$ that a person with a single given name n has gender $\gamma \in \{\text{"Female"}, \text{"Male"}\}$. $D_F(name)$ and $D_M(name)$ are the frequencies of the $name$ in the vocabularies of female D_F and male D_M names. The smoothing variable α prevents the probabilities from getting near-zero values in ambiguous cases. In this way, e.g., names with only a few samples do not affect the final result too much. Likewise, if a name does not appear in either vocabulary, the estimate reduces to 50%—a natural choice for a prior probability when estimating an unknown gender. Equation 2 defines the probability that a given sequence of names $N = (name_1, name_2, \ldots)$ relates to gender γ. To simplify the calculations, the correlation between the names in the sequence was theorized to be statistically independent, e.g., having $name_1$ would not correlate with having $name_2$. Besides, the used vocabularies do not include information about the co-occurrences of given names. Therefore the probability of a sequence could be calculated as a product of the probabilities for each name.

$$\rho(\gamma|name) = \frac{\rho(name|\gamma) \cdot \rho(\gamma)}{\rho(name)} \approx \frac{D_\gamma(name) + \alpha}{D_F(name) + D_M(name) + 2\alpha} \quad (1)$$

[20] `https://nlp.ldf.fi`

[21] `http://nlp.ldf.fi/gender-identification`

$$\rho(\gamma|N = (name_1, name_2, \ldots)) = \frac{\prod_{n \in N} \rho(\gamma|n)}{\prod_{n \in N} \rho(\text{"Female"}|n) + \prod_{n \in N} \rho(\text{"Male"}|n)} \quad (2)$$

For the final decision making, a threshold value τ (e.g., $\tau = 0.75$) is used. For example, if $\rho(\text{"Female"}|N) > \tau$, then the person is classified as a female, or as a male in case $\rho(\text{"Male"}|N) > \tau$. Moreover, no inference is made in the range $\rho \in [1.0-\tau, \tau]$ where the gender remains undefined. For example, when analyzing a unisex name like *Dominique*, the result remains undefined, but adding another name *Gaston*, the application interprets the sequence *Dominique Gaston* as a male name, or as a female in the case *Gabrielle Dominique*.

Person Name Finder Service The Person Name Finder is an API service for identifying references to people and collecting context around them from texts. It utilizes the HENKO ontology to identify person names from texts as a NER task. The Person Name Finder uses the linkage of the family names to places and vocations to differentiate between them and person names. In case the application finds from a text a reference to a single family name and there are no full names with the same family name in the text, it checks if the name is linked to either a place or vocation. If the family name has been linked to a place name, the application returns the place reference to indicate that the name can also be a place. The same procedure is applied to vocations; if a sentence starts with a name that is linked to a vocation written with a capital letter in a beginning of a sentence, the application returns the vocation link. Otherwise, the application returns only person names with links to the person name ontology. In addition, the service can identify information around the name such as times of birth and death, and the gender by utilizing the Gender Identification Service.

The service identifies person names and returns the result set in JSON format. It has been designed to aid in the extraction of personal information from registry entries and natural language texts. The result set contains full names and offers information related to the name such as location in text, links to HENKO, and optionally contextual information, such as gender, dates within brackets, etc. The API and its description[22] are available at the SeCo Text Annotation Service. Currently, the application is being developed and used as a part of named entity recognition and linking to identify person names from the legal and biographical texts. It has been able to identify most names and even some older names, and to enrich them with information such as years within brackets, and gender.

5 Evaluation

This section evaluates the enriching methods for the initial data in Section 2 and the Gender Identification Service from Section 4.

[22] `http://nlp.ldf.fi/api-documentation/#api-NameFinder`

The use of NLP methods for data enrichment provided satisfactory results. The identification of matronymics and patronymics was calculated for 1000 random samples. The F1-score for identification of matronymics was 87.27% and for patronymics 94.42%. Most frequently encountered issue with identification was the lack of Swedish or Russian given names from which the form is derived from. The extraction of suffixes and particles worked well. The F1-score for a sample of 1000 names was 92.78% for suffixes and 100% for particles. The suffix extraction failed for rarer non-Finnish names because they could not be hyphenated correctly due to language identification or lack of hyphenation support.

The linking of names succeeded with varying results. Roughly 23 600 names are linked to Wikidata, and 2500 to DBpedia. The rest of the names could not be linked because either the database did not include the name or there were errors in the data. Often older or less popular names could not be found in either target ontology. Also, some Asian names were linked to several entities in Wikidata with the same label, e.g. *Jin* was linked to two Chinese and one Korean name. The linking of names to topics matched to 785 places and 30 vocations. The success of the linking depended on the quality and coverage of the target ontology. Names from pre-Christian era could not be linked to places or vocations because the target ontologies do not contain a historical vocabulary for the entities.

The Gender Identification Service was evaluated using the names of the relatives extracted from BiographySampo data. It recognized 97.70% of the unique names leaving out only very rare or foreign names. In the test set, all recognized genders were inferred correctly [16]. In addition to using given names, the gender can be concluded e.g. by occupation, by known family relations, or by external contextual information. For example, in the case of AcademySampo all students starting earlier than in 1870 are male [17] since female students were not allowed.

6 Data Service

The person name ontology is published as Linked Open Data on the Linked Data Finland (LDF.fi) platform [8], adhering to the FAIR principles[23]. The platform provides a public SPARQL endpoint[24], IRI dereferencing capabilities, including a generic RDF browsing user interface, and a dataset homepage[25] with general documentation based on the SPARQL Service Description[26], containing a Vocabulary of Interlinked Datasets (VoID) description[27] of the dataset. For human-readable data model documentation[28], we use LODE [27]: when dereferencing IRIs of the name ontology's schema, the user is redirected to a page listing the classes and properties used. The ontology is also published in the ONKI Light

[23] `https://www.go-fair.org/fair-principles/`
[24] `http://ldf.fi/henko/sparql`
[25] `http://ldf.fi/dataset/henko`
[26] `https://www.w3.org/TR/sparql11-service-description/`
[27] `https://www.w3.org/TR/void/`
[28] `http://ldf.fi/schema/henko/`

service[29], where it is searchable and browsable using SKOSMOS[30], a web-based SKOS browser. The data is served on the Apache Jena Fuseki triplestore. The Fuseki runtime and the person name ontology data are built into a Docker image[31] which can be easily rebuilt when there is a need to publish a new version of the data, by simply updating the data in a Git repository.

7 Conclusions

This paper presents the person name ontology HENKO that consists of Finnish person names from the 3rd century to present time. Unlike actor ontologies and vocabularies such as ULAN and BIO, HENKO concentrates on describing person names and their features. The ontology is published as linked open data that connects to AcademySampo, BiographySampo, Norssi Alumni datasets and semantic portals, Wikidata, DBpedia, YSO places, and AMMO ontologies. Its unique data model was influenced by largely used ontologies and vocabularies such as Wikidata, Schema.org, and DBpedia. Out of these ontologies, Wikidata has the most extensive model thus far for names; it divides names by gender, includes etymological information, and has pronunciation instructions. In addition, the Wikidata ontology differentiates patronymic and matronymic names. In contrast, HENKO consists of a large set of Finnish names of which nearly 45% could be linked to Wikidata. In addition, the HENKO has more information about the names such as their usage statistics, linguistic information (suffixes, particles), and provenance information. HENKO model can be used as is for simple patterns consisting of given and family names. In addition, by adding the modelling for middle names, it can be used for wider range of naming conventions. Hence, the ontology is a novel resource for different applications. It can also be used as training material for deep learning based NLP applications alike.

The accuracy of extracting particles and suffixes was satisfactory. The minor issues of suffix extraction could be solved by identifying and splitting family names that are compound words with tools such as the Turku dependency parser [12] or LAS's morphological analyzer. In addition to family names, also given names can contain suffixes that have so far been ignored. They can, e.g., indicate the bearer's gender, like in the female Wilhelmiina based on the male name Wilhelm. The identification and extraction of suffixes enables data analysis for the names. For example, in the history of Finnish last names [26], there have been periods when it has been popular to change Swedish or Russian names to Finnish names with suffixes such as *-la* or *-nen*. When analyzing the AcademySampo data, we found out that family names with suffix *-nen* start to appear only after 1830. To analyze the temporal characters of family names with other suffices remain as future work. Given names [28] have also been modified but by the clergy keeping the parish registries according to the guidelines of different central governments; for example the name Gregorius has been changed to the

[29] `http://light.onki.fi/henko/en/`
[30] `http://skosmos.org`
[31] `https://hub.docker.com/r/secoresearch/fuseki/`

Finnish name Reijo[32]. One future research direction for enriching the data could be to represent there changes of names based on genealogical data and track the changes and suffixes in different linked source datasets. This would also aid in named entity linking (NEL), as the name changes in historical documents could be understood and references to people could be disambiguated better if indicated that the person used different changed names. Modeling of the changes of names has been researched earlier, e.g., in the context of biological taxa [30,3].

The linking of family names to places and vocations enriched the ontology and added context to names. The Person Name Finder utilizes the added context to identify possibly ambiguous nouns when it is used to identify names from text. Unlike typical NEL tools [23,24,5] that concentrate on simply linking entities to knowledge bases, the application can be utilized to extract names from texts and enrich them with contextual information. The Person Name Finder application is still under work, and will be further developed to ease linking to related actor ontologies. In addition to topical linking, in the future, place name linking can be used similarly to, e.g., Tuomas Salste's work[33] by locating the origin of names and visualizing them on a map to aid in genealogical research. By using the extracted suffixes, the linking of names to places could be improved and expanded to names that refer to places but contain a suffix that prevents linking (e.g., Savola referes to Savo without the -la suffix).

The usage statistics of the names enables the usage of the ontology in the Gender Identification Service. Although the functionality of the service is straightforward and based on relative trivial statistics, e.g., it does not consider the co-occurrence of the names and it does not return an estimate for names missing in the ontology, the results have been feasible in our use cases. Related to our service, there are commercial projects such as genderize[34] and gender-api[35] that also use name vocabularies for decision making. Attempt to infer the gender by the ending of the name [1] is problematic with Finnish names where, e.g., *Jari* and *Kari* are male names but *Sari* and *Mari* female ones. A blog post [2] by Ellis Brown introduces a project where the gender is inferred from character sequences in names using a recurrent neural network. Due to the feasible results for our use cases, we have not implemented similar algorithms for inferring the gender for names missing from our vocabulary.

Acknowledgments This work is part of the Anoppi project[36] funded by the Ministry of Justice in Finland. Thanks to Aki Hietanen, Saara Packalén, Tiina Husso, and Oili Salminen of the Ministry of Justice, and Risto Talo, Jari Linhala, and Arttu Oksanen of Edita Publishing Ltd. for collaboration. Thanks also to Aleksandra Konovalova from University of Helsinki and Esko Kirjalainen from The Finnish Digital Agency for insightful discussions. CSC – IT Center for Science, Finland, provided us with computational resources.

[32] `https://www.genealogia.fi/nimet/nimi15s.htm`
[33] `https://www.tuomas.salste.net/suku/nimi/`
[34] `https://genderize.io`
[35] `https://gender-api.com`
[36] `https://seco.cs.aalto.fi/projects/anoppi/en/`

References

1. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit, chap. 6.1.1. O'Reilly Media, Inc. (2009)
2. Brown, E.: Gender Inference from Character Sequences in Multinational First Names. https://towardsdatascience.com/name2gender-introduction-626d89378fb0, accessed: 2020 Mar 3
3. Chawuthai, R., Takeda, H., Wuwongse, V., Jinbo, U.: Presenting and Preserving the Change in Taxonomic Knowledge for Linked Data. Semantic Web **7**(6), 589–616 (2016)
4. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing Wikidata to the linked data web. In: International Semantic Web Conference. pp. 50–65. Springer (2014)
5. Francis-Landau, M., Durrett, G., Klein, D.: Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks. arXiv preprint arXiv:1604.00734 (2016)
6. Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: Evolution of Structured Data on the Web. Communications of the ACM **59**(2), 44–51 (2016)
7. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with Wikipedia. Artificial Intelligence **194**, 130–150 (Jan 2013). https://doi.org/10.1016/j.artint.2012.04.005, http://dx.doi.org/10.1016/j.artint.2012.04.005
8. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. In: The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers. pp. 226–230. Springer-Verlag (May 2014)
9. Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., Sirola, L.: Reassembling and Enriching the Life Stories in Printed Biographical Registers: Norssi High School Alumni on the Semantic Web. In: Proceedings, Language, Technology and Knowledge (LDK 2017). pp. 113–119. Springer-Verlag (June 2017)
10. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K.: Linked Data – A Paradigm Change for Publishing and Using Biography Collections on the Semantic Web. In: Proceedings of the Third Conference on Biographical Data in a Digital World (BD 2019) (September 2019)
11. Joudrey, D., Taylor, A., Miller, D.: Introduction to Cataloging and Classification. Libraries Unlimited, 11 edn. (2015)
12. Kanerva, J., Ginter, F., Miekka, N., Leino, A., Salakoski, T.: Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics (2018)
13. Klimek, B.: Proposing an OntoLex-MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models. In: LDK Workshops. pp. 68–73 (2017)
14. Klimek, B., Arndt, N., Krause, S., Arndt, T.: Creating Linked Data Morphological Language Resources with MMoOn – The Hebrew Morpheme Inventory. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 892–899 (2016)
15. Koho, M., Gasbarra, L., Tuominen, J., Rantala, H., Jokipii, I., Hyvönen, E.: AMMO Ontology of Finnish Historical Occupations. In: Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19). vol. 2375, pp. 91–96. CEUR Workshop Proceedings (June 2019), vol 2375

16. Leskinen, P., Hyvönen, E.: Extracting Genealogical Networks of Linked Data from Biographical Texts. In: Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019), Portoroz, Posters & Demonstrations (June 2019)
17. Leskinen, P., Hyvönen, E.: Linked Open Data Service about Historical Finnish Academic People in 1640–1899. In: Proceedings of Digital Humanities in Nordic Countries (DHN 2020), Riga. CEUR Workshop Proceedings (March 2020)
18. Mäkelä, E.: Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text. In: European Semantic Web Conference. pp. 424–428. Springer (2014)
19. Mäkelä, E.: LAS: an integrated language analysis tool for multiple languages. The Journal of Open Source Software **1**(6) (October 2016)
20. Malmi, E., Rasa, M., Gionis, A.: AncestryAI: A Tool for Exploring Computationally Inferred Family Trees. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 257–261. International World Wide Web Conferences Steering Committee (2017)
21. McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., Cimiano, P.: The OntoLex-Lemon model: development and applications. In: Proceedings of eLex 2017 conference. pp. 19–21 (2017)
22. McCrae, J.P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., De Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., et al.: The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 2435–2441 (2016)
23. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th international conference on semantic systems. pp. 1–8. ACM (2011)
24. Nguyen, D.B., Hoffart, J., Theobald, M., Weikum, G.: AIDA-light: High-throughput named-entity disambiguation. In: Proceedings of the Workshop on Linked Data on the Web (LDOW 2014), co-located with the 23rd International World Wide Web Conference (WWW 2014). vol. 1184. CEUR Workshop Proceedings (April 2014)
25. Oksanen, A., Tamper, M., Tuominen, J., Hietanen, A., Hyvönen, E.: Anoppi: A pseudonymization service for Finnish court documents. In: Araszkiewicz, M., Rodríguez-Doncel, V. (eds.) Legal Knowledge and Information Systems. JURIX 2019: The Thirty-second Annual Conference. pp. 251–254. IOS Press (December 2019)
26. Paikkala, S.: Sukunimet sukututkimuksessa. Sukutieto: Sukutietotekniikka ry:n jäsenlehti **14**(4) (1997)
27. Peroni, S., Shotton, D., Vitali, F.: Tools for the Automatic Generation of Ontology Documentation: A Task-Based Evaluation. International journal on Semantic Web and information systems **9**(1), 21–44 (2013)
28. Rajasuu, R.: Kuopiossa, Oulussa ja Turussa vuosina 1725–1744 ja 1825–1844 syntyneiden kastenimet. Ph.D. thesis, University of Eastern Finland (2013)
29. Tamper, M., Hyvönen, E., Leskinen, P.: Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research. In: Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICling 2019). Springer-Verlag (April 2019)
30. Tuominen, J., Laurenne, N., Hyvönen, E.: Biological Names and Taxonomies on the Semantic Web – Managing the Change in Scientific Conception. In: Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011). pp. 255–269. Springer-Verlag (June 2011)