The RICHFIELDS Framework for Semantic Interoperability of Food Information Across Heterogenous Information Systems

Tome Eftimov[1], Gordana Ispirova[1,2], Peter Korošec[1,3] and Barbara Koroušić Seljak[1]
[1]*Computer Systems Department, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia*
[2]*Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia*
[3]*Faculty of Mathematics, Natural Sciences and Information Technologies, Glagoljaška ulica 8, 6000 Koper, Slovenia*

Keywords: Semantic Interoperability, RICHFIELDS Ontology, Food Information, Ontology Population, Semantic Annotation.

Abstract: In an EU-funded project RICHFIELDS, a data platform was designed with the aim to collect, link and harmonize, analyze, store, and deliver food- and nutrition-related data and information to various stakeholders. To integrate heterogenous food data sets, we propose a RICHFIELDS framework for semantic interoperability of food information, which is a combination of already developed NLP approaches for the food domain. The framework includes i) a food ontology to which foods are linked, ii) a part that explains how the relevant foods can be extracted and represented in a structured way, and iii) a similarity measure that is used to link the foods to the ontology. To evaluate the RICHFIELDS framework, we selected two distinct data sets from different food information systems. The experimental results provided promising results,i.e., 81.5% and 87.5% of the foods from the first and the second data set, respectively, obtained a tag from the ontology (i.e., semantic annotation was performed). The annotations provided by the framework allow automatic integration of food information provided in both data sets.

1 INTRODUCTION

Creating a healthy diet requires a lot of information and knowledge from food science. Nowadays, there are many information systems that provide food- and nutrition- related data. These systems can be either: a scientific cloud (e.g., European Open Science Cloud (EOSC, 2018), Zenodo (Zenodo, 2018), and FigShare (FigShare, 2018)), a server (e.g., Quisper (QuaLiFY, 2018), EuroFIR (EuroFIR, 2018), and GS1 GDSN (Global Data Synchronisation Network) (GDSN, 2018)), or application (e.g., PRECIOUS (PRECIOUS, 2018), FitBit, Twitter, and Facebook). Each system uses its own way of describing information. In order to exchange data with unambiguous, shared meaning, semantic interoperability is a requirement to enable machine computable logic, inferencing, knowledge discovery, and data federation between information systems. This is made by adding metadata about the data, linking each data element to an ontology (i.e., semantic data model). For this reason, in autumn 2015, the H2020 project RICHFIELDS started with the aim is to collect, link and harmonize, analyze, store and deliver food- and nutrition-related data and information to various stakeholders. Data may be of any type, i.e., structured, semi-structured or unstructured; small or big; open or linked, raw or aggregated. To make this possible, semantic enrichment should be applied to solve some of the most common problems, allowing for: effective search in databases, integration of heterogeneous data sets, faster information retrieval, regularly updated domain knowledge, etc. Since semantic enrichment involves adding metadata to the data, or linking specific data to an ontology, in the case of RICHFIELDS, a domain ontology that covers food- and nutrition-related domain should be specified. Having such a representation of the domain, there are several questions that must be addressed: what type of data (e.g., structured, semi-structured, or unstructured) needs to be harmonized; if the data is unstructured, how we can extract the relevant data that should be linked to the domain ontology; what is the similarity measure that will be used for linking data to the ontology; etc. To address such questions in the case of RICHFIELDS, we propose a framework for semantic interoperability of food information across heterogenous information systems.

Eftimov, T., Ispirova, G., Korošec, P. and Seljak, B.
The RICHFIELDS Framework for Semantic Interoperability of Food Information Across Heterogenous Information Systems.
DOI: 10.5220/0006951703150322
In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 1: KDIR*, pages 315-322
ISBN: 978-989-758-330-8

The paper is organized as follows: Section 2 gives an overview of the related work. Section 3 introduces the RICHFIELDS framework for semantic interoperability of food information. Section 4 presents a RICHFIELDS case study, in which discussion of results is presented. The conclusions are presented in Section 5.

2 RELATED WORK

To make ambiguous and heterogenous content interlinked, semantic interoperability plays an important role. It involves two steps: selecting or developing an ontology that describes the domain, and enriching relevant data with metadata, which are machine-processable data pieces (i.e., tags from the ontology). However there are subquestions related to each step that need to be answered.

Having a domain-ontology that describes the domain in general is a challenging task because each ontology is developed for a specific application scenario. In the domain of food, several food ontologies already exist, such as: FoodWiki, AGROVOC, Open Food Facts, Food Product Ontology, Foods, and FoodOn. A detailed review of the aforementioned food ontologies is provided in (Boulos et al., 2015). The problem of generating a general food domain ontology has been partly solved by the QuaLiFY European project (http://quisper.eu/), where existing food information systems were explored by scientific bodies like EuroFIR (European Food Information Resource Network) and NuGO (http://www.nugo.org/).

After a domain ontology is selected, the relevant data should be linked to the ontology. The questions that appear here are: how can we extract the relevant data (e.g., especially if the data is unstructured), and how the extracted data can be linked to the ontology in an automatic way. If the data is represented as structured or semi-structured, different rule-based approaches could be applied in order to extract the relevant data that further will be linked to the ontology. On the other hand, when the data is unstructured (i.e., represented as text), a more complex scenario is presented, in which information extraction (IE) methods should be applied in order to extract the relevant data. Nowadays, the IE from the biomedical literature is a very important task in order to improve public health. IE is a task of automatically extracting information from unstructured data and in most cases concerns processing of human language texts by means of natural language processing (NLP) (Aggarwal and Zhai, 2012; Piskorski and Yangarber, 2013). The information to be extracted is predefined by users, and consists of predefined concepts of interest (entities), relationships between them and events. One of the classic IE tasks is named-entity recognition (NER), which addresses the problem of the identification and classification of predefined concepts (entities). Various NER methods exist: terminology-driven NER methods (Miller et al., 1992; Aronson, 2001; Zhou et al., 2006), rule-based NER methods (Farmakiotou et al., 2000; Petasis et al., 2001; Hanisch et al., 2005), corpus-based NER methods (Rindflesch et al., 2000; Rocktäschel et al., 2012; Alnazzawi et al., 2015; Leaman et al., 2015), NERs based on active learning (Settles, 2010), and NERs that use deep neural networks (Collobert and Weston, 2008; Collobert et al., 2011; Chiu and Nichols, 2015; Huang et al., 2015; Santos and Guimaraes, 2015; Lample et al., 2016; Habibi et al., 2017; Lopez and Kalita, 2017). Because NER methods with best performances are usually corpus-based NER methods, there is a need for annotated corpus from biomedical literature that will include the entities of interest. For this purpose, different annotated corpora are produced by shared tasks, where the main aim is to challenge and encourage research teams on NLP problems.

To allow automatic integration of extracted information from a NER task, the information needs to be further processed. The problem that appears is that the same entity can be mentioned in different ways in the same or different documents, using different phrases regarding the text variability. To collect the information for a given entity or even more to combine the information for the entity from different documents, it is crucial to map the entity to a concept that exists in a terminological resource (i.e., an ontology). By mapping it to a concept from a terminological resource, the extracted entity receives a unique identifier which is the identifier for that entity in the terminological resource. Having unique identifiers helps the process of collecting and combing the information for some entity, even if it has different textual representations. The process of automatic mapping between an entity in text and a concept in a terminological resource is known as text normalization. Many normalization methods are based on string similarity measures. String similarity measures give us a metric for similarity (or dissimilarity) between two text strings (Metzler et al., 2007; Gomaa and Fahmy, 2013). They can be performed on a character, term level, or a mix of both. Also, there are methods that use post-processing rules applied after the concept is extracted or regular expressions to find some matches for which a specific form may not occur in the terminological resource (Ramanan et al., 2013). Some normalization methods are based on ranking techni-

que in order to rank the candidate matches and then to find the most relevant one (Collier et al., 2015). An example of automatic normalization method focussed on phenotypic information, which integrates a number of different similarity measures, is presented in (Alnazzawi et al., 2016). Normalization methods can also use ML algorithms to improve results, which was shown in the gene normalization task as part of BioCreative II (Morgan et al., 2008) and BioCreative III (Lu et al., 2011). To the best of our knowledge, no previously reported automatic normalization method has focussed on the food domain.

3 THE RICHFIELDS FRAMEWORK FOR SEMANTIC INTEROPERABILITY

Let us assume that food information systems share data sets that should be integrated. To make them understandable in machine-readable format, the RICHFIELDS framework for semantic interoperability is proposed. The framework is presented in Figure 1 and consists of four steps:

1. Select the domain ontology;
2. Apply pre-processing of the data with regard to its type to obtain structured data;
3. Link the structured data to concepts that exist in the ontology using a similarity measure;
4. Perform semantic annotation or ontology population.

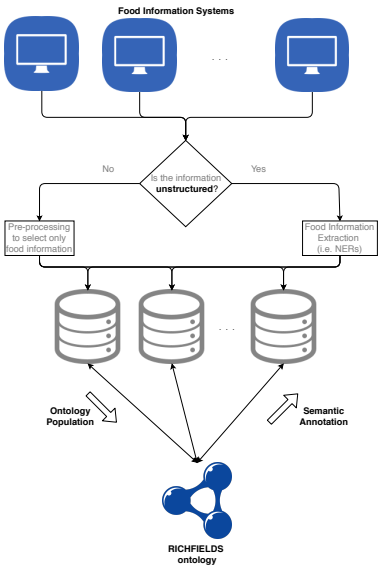

Figure 1: The RICHFIELDS framework for semantic interoperability.

The first step is to select the domain ontology that is a model to which the data will be linked and from which the metadata will be used to make the data understandable in machine readable format.

Next, the type of the data should be defined i.e. structured, semi-structured, or unstructured. If we are dealing with structured or semi-structured data, preprocessing can be made by applying some heuristics (e.g., rules based on regular expressions) in order to select the data that will be linked. If the data is unstructured, first information extraction methods (i.e. named-entity recognition methods, NERs) should be applied to extract and structure the relevant information that will be linked to the ontology.

The third step defines a similarity measure used for linking the data to concepts that already exist in the ontology. Since most of the data is presented as text, different text normalization methods that involve text similarity measures can be applied.

Finally, according to the value of the similarity measures, semantic annotation or ontology population should be performed. A threshold value for the similarity measure needs to be defined. If the value of the similarity measure is greater or equal than the selected threshold, a semantic annotation should be performed. This means that this matching is a good one, the concept of searching already exists in the ontology, so it should be annotated with the metadata. In this case, the data set is changed by including tags from the ontology. In the other case, when the value of the similarity measure is lower than the selected threshold, the data cannot be linked to the ontology because such concept does not exist in the ontology, so ontology population should be performed. This involves adding an instance for such concept in the ontology.

In general, the above mentioned steps are familiar for each framework used for semantic interoperability, the questions that appear are related to information extraction methods and the definition of the similarity measure used for linking, since each domain is specific and it follows that if some methods are good for a specific domain it does not follow that they will also be good for other domains.

4 RICHFIELDS CASE STUDY

We continue by explaining in detail each part of the RICHFIELDS framework for semantic interoperability. First, the food ontology is explained, followed by the methods that can be used for food information extraction from unstructured data. Then, the similarity measure used for linking foods to the RICHFIELDS

ontology is reintroduced. Finally, the results of linking two food-related data to ontology are explained and discussed.

4.1 The RICHFIELDS Ontology

The development of the ontology that is used by RICHFIELDS started from the Quisper ontology, which was previously developed by JSI as part of the EU-funded project QuaLiFY (QuaLiFY, 2018). The ontology consists of six super classes: Component, Food, FoodGroup, Personal, Single Nucleotide Polymorphism, and Unit, which were further described with data and object properties. RICHFIELDS covers a wider domain than the QuaLiFY project and for this reason, the Unit concept was replaced with a concept with the same name from a widely-used ontology, called Units of Measurements Ontology (UO) (Gkoutos et al., 2012), which is currently being used in many scientific resources for the standardized description of measurements units. Also, a new concept, MatrixUnit, was added with the corresponding subconcepts for matrix units that can be found in the EuroFIR Thesauri. Because this study was focused on semantic interoperability of food-related data sets, the RICHFIELDS ontology was updated by populating the Food concept with 5,416 food individuals from FoodEx2 data (EFSA, 2017), which were also described with two data properties FoodName and FoodEx2 code that are specific for the FoodEx2 representation.

4.2 drNER

If we are working with unstructured data (i.e., represented as text) the first step is to extract the relevant food data that should be linked to the ontology. For this reason, NERs should be applied. From an overview of the existing IE methods from the biomedical literature, a lot of NER methods that exist in the domain of biomedical literature are focused on different biomedical domains. The commonly used NER methods are the corpus-based NER methods that rely on annotated corpus for the domain of interest, which is produced by the domain experts. Several studies are conducted in the dietary domain, but with different goals. For example, Xia et al. (Xia et al., 2013) presented an approach to identify rice protein resistant to *Xanthomonas oryzae pv. oryzae*, which is an approach to enhance gene prioritization by combining text mining technologies with a sequence-based approach. Co-occurrence methods were also used to identify ingredients mentioned in food labels and extracting food-chemical and food-disease relationship (Müller et al., 2004; do Nascimento et al., 2013; Jensen et al., 2014). We did not find any research that focuses on extracting dietary information from evidence-based dietary recommendations and for this reason we recently proposed a rule-based NER method, known as drNER (Eftimov et al., 2016; Eftimov et al., 2017c). It is a combination of a terminological-driven NER and rule-based NER. The difference with purely terminological-driven NERs is that we do not only use dictionaries with concepts and synonyms (as terminological resources), but we allow the reuse of some corpus-based NERs that exist for some entities. If corpus-based NERs exist for some entities we are interested in, we use them to annotate text data and then to see if some tokens have labels that correspond to entities of interest. We also combine corpus-based NERs that exist for some entities in which we are interested, following the idea of ensemble learning in order to achieve better performance than the performance obtained from any corpus-based NER alone. The difference with the rule-based NERs is that we do not use rules associated with the characteristics of the entities. This is because having rules for each of the entities we are interested in requires too much time and effort to produce them. We only used a small number of Boolean algebra rules that are not related to the characteristics of the entities, but help us define the phrases that are the entities mentions. Evaluation of the method showed that the method gives promising results and can be used for information extraction of evidence-based dietary recommendations.

4.3 StandFood

When the foods are represented in structured way, the next step is to link each of them to a concept that has already existed in the ontology. To do this, text normalization methods that are based on string similarity measures should be applied. In (Eftimov et al., 2017a; Eftimov et al., 2017b), we presented a method, known as StandFood, which is used for standardization of foods according to FoodEx2 that is a comprehensive food classification and description system for exposure assessment introduced by EFSA (EFSA, 2017). StandFood is a semi-automatic system for classifying and describing foods and consists of three parts: the first classifies the food concept into one from four FoodEx2 categories (i.e., raw, derivatives, simple composite, and aggregated composite) using ensemble of classifiers, the second describes the food concept using the FoodEx2 code using natural language processing approach, and the third combines the results from the first and the second part to improve the result for the classification part by defining post-processing rules. As a similarity measure in

the RICHFIELDS framework, we use the description part of StandFood. For this reason, we are going to reintroduce this part that uses the POS-tagging probability weighted method (Eftimov and Koroušić Seljak, 2015).

For the similarity measure, let D_1 and D_2 are the names of two foods that are linked. Let us define

$$\begin{aligned} N_i &= \{nouns\ extracted\ from\ D_i\}, \\ A_i &= \{adjectives\ extracted\ from\ D_i\}, \\ V_i &= \{verbs\ extracted\ from\ D_i\}, \end{aligned} \tag{1}$$

where $i = 1,2$. To find the similarity between these two food names, an event is defined as a product of two other events

$$X = N(A+V), \tag{2}$$

where N is the similarity between the nouns found in N_1 and N_2, and $A+V$ is the similarity between the two sets of adjectives and verbs handled together as A_1+V_1 and A_2+V_2. The adjectives and verbs are handled together to avoid different forms with the same meaning. For example, if adjectives and verbs are handled separately, the match "apple dry" and "dried apples" will not be a perfect match. To avoid this, lemmatization is applied for each extracted noun, verb and adjective, and the similarity event uses their lemmas. Because these two events are independent, the probability of the event X can be calculated as

$$P(X) = P(N)P(A+V). \tag{3}$$

For this, the probabilities of each of the two events need to be defined. Because the problem looks for the similarity between the two sets, it is logical to use the Jaccard index, J, which is used in statistics for comparing similarity and diversity of sample sets (Real and J.M.Vargas, 1996). For the similarity between the nouns, the Jaccard index is used, while for the similarity between the adjectives and verbs the Jaccard index is used in combination with Laplace probability estimate (Cestnik et al., 1990), this is because in some food names the additional information provided by the adjectives or verbs can be missed, but the relevant match can be found, so there will be no zero probabilities. The probabilities are calculated as

$$\begin{aligned} P(N) &= \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|}, \\ P(A+V) &= \frac{|(A_1 \cup V_1) \cap (A_2 \cup N_2)| + 1}{|(A_1 \cup V_1) \cup (A_2 \cup N_2)| + 2}. \end{aligned} \tag{4}$$

By substituting Equation 4 into Equation 3, we obtain a weight for each matching pair. Finally, the pair with the highest weight is the most relevant found match. More details about the POS-tagging probability weighted method can be found in (Eftimov and Koroušić Seljak, 2015).

In the case of RICHFIELDS, each preprocessed data set that contains foods is linked to the RICHFIELDS ontology in a way that each food concept is linked to food individuals that exist in RICHFIELDS ontology using the POS-tagging probability weighted method. At the end the pair with the highest value of the similarity measure is selected as the real one. However, it can happen that non returned match is true. One reason for this could be that such food concept does not exist as a food individual in the RICHFIELDS ontology. For this reason the similarity measure value that is returned as a match is further checked with a threshold value given as a priori information, which in the case of RICHFIELDS is set at 0.125 and it comes from experimental evaluations performed on food matching problem. If the value is greater or equal than 0.125 then the food concept in the data set is annotated using the tag for the food individual from the ontology, otherwise we cannot find a match, the concept does not exist as individual in the ontology, so the RCIHFIELDS ontology must be populated with this concept.

4.4 Food Information Systems

To show how the RICHFIELDS framework for semantic interoperability works, we used two food-related data sets that are provided from two food information systems (i.e., PRECIOUS and GS1 GDSN) that rely on different standards, which are related to the same concepts but use different terminology and classification. PRECIOUS and GS1 GDSN provide data in semi-structured form (i.e., JSON format and GS1 XML format, respectively).

4.4.1 PRECIOUS

PRECIOUS is a mobile app for preventive health and wellbeing care that was developed in the FP7 project PREventive Care Infrastructure based On Ubiquitous Sensing (PRECIOUS, 2018). It was decided to collect different kinds of biometric data (e.g. nutrition, physical activity, sleep, etc.). Our PRECIOUS data set consists of 437 foods, some of them are described in English and some in Spanish. An example of one food concept from the PRECIOUS data set is given in Figure 2.

4.4.2 GS1 GDSN

The GS1 Global Data Synchronisation Network is a network of interoperable data pools enabling collaborating users to securely synchronise master data

```
{
        _id: ObjectId(**),
        key: FOOD_INTAKE,
        from: ISODate(2017-09-01T21:54:39.463Z),
        to: ISODate(2017-09-01T21:54:39.464Z),
        value: [{
                food_amount: 80,
                food_name: Nectarina,
                food_type: 5,
                        }],
        id: 1470088479463,
        user: ObjectId(ffffffffffffffffffffffff),
        posted: ISODate(2017-09-02T09:06:33.484Z),
        __v: 0
}
```

Figure 2: An example of food concept from the PRECIOUS data set.

based on GS1 standards. GDSN supports accurate, real-time data sharing and trade item updates among subscribed trading partners. Currently available GDSN standards for nutrition and health are available at https://www.gs1.org/gdsn-standards. The data provided by GS1 consists of 25 foods provided by the GS1 Slovenia. All foods are available with their Slovenian and English names.

4.5 Results

Since the data from PRECIOUS data set is semi-structured, first, by using regular expressions we parsed the document to structure it. Then, we split the data set into two parts: Spanish names and English names. The Spanish names are translated in English using a simulation that involves scraping with Selenium and Google Translate web site. After that, the translated Spanish food names are merged with the existing English food names, which results in one data set that will be linked to the RICHFIELDS ontology. For the GS1 GDSN data we used the English names that are provided when we linked it to the ontology.

In the process of linking 81.5% of the foods from the PRECIOUS data set obtained tags from the RICHFIELDS ontology, while we need to populated the ontology for the other 18.5% when the match does not exist. In the case of GS1 GDSN, 87.5% of the foods obtained their tags from the ontology, while 12.5% were included as new food individuals.

An annotated example from the PRECIOUS data set is presented in Figure 3, in which the food concept is described by an additional RICHFIELDS tag that is the tag for the same food concept that exists in the ontology.

Annotated examples from the GS1 GDSN are presented in Table 1. The Global Trade Item Number (GTIN) can be used by a company to uniquely identify all of its trade items. In our study, we presented it as a string (e.g., GTIN1), in order not to invade the privacy of the real data.

```
{
        _id: ObjectId(**),
        key: FOOD_INTAKE,
        from: ISODate(2017-09-01T21:54:39.463Z),
        to: ISODate(2017-09-01T21:54:39.464Z),
        value: [{
                food_amount: 80,
                food_name: Nectarina,
                food_type: 5,
                RICHFIELDS_tag:http://www.semanticweb.org/
tome/ontologies/2018/2/Richfields#A01GN"
        }],
        id: 1470088479463,
        user: ObjectId(ffffffffffffffffffffffff),
        posted: ISODate(2017-09-02T09:06:33.484Z),
        __v: 0
}
```

Figure 3: An example of annotated food concept from the PRECIOUS data set.

Though each food individual is described with a FoodName and a FoodEx2 code, in the case of ontology population we cannot provide the FoodEx2 code for the new individual. Ontology population happens when the food individual does not exist in the ontology, which means that it does not exist in the FoodEx2 data set since the ontology is populated with all existing foods from the FoodEx2 data set. From here, a new problem arises, which is how to generate a FoodEx2 code for new food individual. This is also one direction for our future work.

5 CONCLUSIONS

To allow integration of heterogenous food data sets, faster information retrieval, and regularly updated food knowledge, we propose a RICHFIEDLS framework for semantic interoperability of food information. The framework includes a food ontology that is the resource to which data sets, which used different standards to describe foods, are linked. Depending on the data type (i.e., structured, semi-structured, or unstructured), pre-processing should be applied to select and represent only food information in a structure way. Then, each food concept is linked to the ontology using a similarity measure. Depending on the similarity measure value, semantic annotation or ontology population should be applied.

To show how the proposed framework works, we used two food-related data sets that are provided from two different food information systems, PRECIOUS and GS1 GDSN. The experiment results provided promising results, where 81.5% and 87.5% of the foods from PRECIOUS and GS1 GDSN obtained a tag from the ontology (i.e., semantic annotation was performed), respectively. Further, the RICHFIELDS ontology annotations allow automatic integration of food information provided in these two data sets.

Food items, for which the linking does not give good results (i.e., the food item does not exist in the

Table 1: Annotated examples of foods from the GS1 data set.

GTIN	Food name (in Slovenian)	English food name	RICHFIELDS tag
GTIN1	Liker z limono	Liqueur with lemonn	http://www.semanticweb.org/tome/ontologies/2018/2/Richfields#A03NS
GTIN2	Kefir z borovnicami	Kefir with blueberries	http://www.semanticweb.org/tome/ontologies/2018/2/Richfields#A02NV
GTIN3	Sirup z malinami	Juice with raspberries	http://www.semanticweb.org/tome/ontologies/2018/2/Richfields#A03CD

ontology), were used for ontology population. However, there is an additional challenge that arises while performing this process, which is also one direction of our future work. The problem is how to generate the FoodEx2 code for a food that does not exist in the ontology, which can often happen when we are working with composite foods (i.e., recipes).

ACKNOWLEDGEMENTS

This work was supported by the project from the Slovenian Research Agency (research core funding No. P2-0098), from the European Union's Horizon 2020 research and innovation program under grant agreement No. 654280 (RICHFIELDS), and from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 621329 (ISOFOOD). We would also like to thank the PRECIOUS team from Aalto University and GS1 Slovenia for providing the data sets that are used in this case study.

REFERENCES

Aggarwal, C. C. and Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Alnazzawi, N., Thompson, P., and Ananiadou, S. (2016). Mapping phenotypic information in heterogeneous textual sources to a domain-specific terminological resource. *PloS one*, 11(9):e0162287.

Alnazzawi, N., Thompson, P., Batista-Navarro, R., and Ananiadou, S. (2015). Using text mining techniques to extract phenotypic information from the phenochf corpus. *BMC medical informatics and decision making*, 15(2):1.

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Boulos, M. N. K., Yassine, A., Shirmohammadi, S., Namahoot, C. S., and Brückner, M. (2015). Towards an internet of food: Food ontologies for the internet of things. *Future Internet*, 7(4):372–392.

Cestnik, B. et al. (1990). Estimating probabilities: a crucial task in machine learning. In *ECAI*, volume 90, pages 147–149.

Chiu, J. P. and Nichols, E. (2015). Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*.

Collier, N., Oellrich, A., and Groza, T. (2015). Concept selection for phenotypes and diseases using learn to rank. *Journal of biomedical semantics*, 6(1):24.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

do Nascimento, A. B., Fiates, G. M. R., dos Anjos, A., and Teixeira, E. (2013). Analysis of ingredient lists of commercially available gluten-free and gluten-containing food products using the text mining technique. *International journal of food sciences and nutrition*, 64(2):217–222.

EFSA ((accessed February 17, 2017)). *The food classification and description system FoodEx2 (revision 2)*. https://www.efsa.europa.eu/.

Eftimov, T., Ispirova, G., Korošec, P., and Koroušić Seljak, B. (2017a). A semi-automatic system for classifying and describing foods according to FoodEx2. In *3rd IMEKO FOODS, Metrology promoting Standardization and Harmonization in Food and Nutrition*, pages 56–59.

Eftimov, T., Korošec, P., and Koroušić Seljak, B. (2017b). Standfood: Standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2. *Nutrients*, 9(6):542.

Eftimov, T. and Koroušić Seljak, B. (2015). POS tagging-probability weighted method for matching the internet recipe ingredients with food composition data. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on*, volume 1, pages 330–336. IEEE.

Eftimov, T., Koroušić Seljak, B., and Korošec, P. (2017c). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PloS One*, 12(6):e0179488.

Eftimov, T., Koroušić Seljak, B., and Korošec, P. (2016). Grammar and dictionary based named-entity linking for knowledge extraction of evidence-based dietary recommendations. In *Proceedings of the 8th international Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, (IC3K 2016)*, volume 1:KDIR, pages 150–157.

EOSC (2018). *European Open Science Cloud*. accessed June 12, 2018.

EuroFIR (2018). *European Food Information Resource*. accessed September 18, 2016.

Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C. D., and Stamatopoulos, P. (2000). Rule-based named entity recognition for greek financial texts. In *Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)*, pages 75–78. Citeseer.

FigShare (2018). *Simplifying your research workflow*. accessed June 12, 2018.

GDSN, G. (2018). *The Global Data Synchronisation Network*. accessed June 12, 2018.

Gkoutos, G. V., Schofield, P. N., and Hoehndorf, R. (2012). The units ontology: a tool for integrating units of measurement in science. *Database*, 2012:bas033.

Gomaa, W. H. and Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13).

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., and Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Hanisch, D., Fundel, K., Mevissen, H.-T., Zimmer, R., and Fluck, J. (2005). Prominer: rule-based protein and gene entity recognition. *BMC bioinformatics*, 6(1):S14.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Jensen, K., Panagiotou, G., and Kouskoumvekaki, I. (2014). Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level. *PLoS computational biology*, 10(1):e1003432.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Leaman, R., Wei, C.-H., Zou, C., and Lu, Z. (2015). Mining patents with tmchem, gnormplus and an ensemble of open systems. In *Proce. The fifth BioCreative challenge evaluation workshop*, pages 140–146.

Lopez, M. M. and Kalita, J. (2017). Deep learning applied to nlp. *arXiv preprint arXiv:1703.03091*.

Lu, Z., Kao, H.-Y., Wei, C.-H., Huang, M., Liu, J., Kuo, C.-J., Hsu, C.-N., Tsai, R. T.-H., Dai, H.-J., Okazaki, N., et al. (2011). The gene normalization task in biocreative iii. *BMC bioinformatics*, 12(8):S2.

Metzler, D., Dumais, S., and Meek, C. (2007). Similarity measures for short segments of text. In *European Conference on Information Retrieval*, pages 16–27. Springer.

Miller, R. A., Gieszczykiewicz, F. M., Vries, J. K., and Cooper, G. F. (1992). Chartline: providing bibliographic references relevant to patient charts using the umls metathesaurus knowledge sources. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 86. American Medical Informatics Association.

Morgan, A. A., Lu, Z., Wang, X., Cohen, A. M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., et al. (2008). Overview of biocreative ii gene normalization. *Genome biology*, 9(2):S3.

Müller, H.-M., Kenny, E. E., and Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS biology*, 2(11):e309.

Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V., and Spyropoulos, C. D. (2001). Using machine learning to maintain rule-based named-entity recognition and classification systems. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 426–433. Association for Computational Linguistics.

Piskorski, J. and Yangarber, R. (2013). Information extraction: past, present and future. In *Multi-source, multilingual information extraction and summarization*, pages 23–49. Springer.

PRECIOUS (2018). *Preventive Care Infrastructure based On Ubiquitous Sensing*. accessed June 12, 2018.

QuaLiFY (2018). *Information service for personalised nutrition and lifestyle advice*. accessed June 12, 2018.

Ramanan, S., Broido, S., and Nathan, P. S. (2013). Performance of a multi-class biomedical tagger on clinical records. In *CLEF (Working Notes)*.

Real, R. and J.M.Vargas (1996). The probabilistic basis of jaccard's index of similarity. *Systematic biology*, pages 380–385.

Rindflesch, T. C., Tanabe, L., Weinstein, J. N., and Hunter, L. (2000). Edgar: extraction of drugs, genes and relations from the biomedical literature. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 517. NIH Public Access.

Rocktäschel, T., Weidlich, M., and Leser, U. (2012). Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.

Santos, C. N. d. and Guimaraes, V. (2015). Boosting named entity recognition with neural character embeddings. *arXiv preprint arXiv:1505.05008*.

Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.

Xia, J., Zhang, X., Yuan, D., Chen, L., Webster, J., and Fang, A. C. (2013). Gene prioritization of resistant rice gene against xanthomas oryzae pv. oryzae by using text mining technologies. *BioMed research international*, 2013.

Zenodo (2018). *Zenodo*. accessed June 12, 2018.

Zhou, X., Zhang, X., and Hu, X. (2006). Maxmatcher: Biological concept extraction using approximate dictionary lookup. In *Pacific Rim International Conference on Artificial Intelligence*, pages 1145–1149. Springer.