Using Linked Data for better navigation in summaries of product characteristics*

Jakub Kozák[1], Martin Nečaský[1], Jan Dědek[2], Jakub Klímek[1], and Jaroslav Pokorný[1]

[1] Charles University in Prague, Faculty of Mathematics and Physics, Czech Republic
{kozak,necasky,klimek,pokorny}@ksi.mff.cuni.cz
[2] Collite Systems, Prague, Czech Republic
jan.dedek@collite.cz

Abstract. Summaries of product characteristics (SPCs) serve as a basic source of information for physicians about medicinal products. SPC is attached to every single registered medicinal product and contains large amount of valuable data in natural language. In this paper we deal with natural language processing (NLP), annotation and Linked Data representation of SPCs. Moreover, we also use the annotations for acquisition of new information about, e.g., interactions. A web application for browsing Linked Data representation of SPCs has been developed.

Keywords: Linked Data, e-health, SPC, NLP, RDFa

1 Introduction

Information about drugs and medicinal products is scattered across various data sources and it is hard for a physician to gather it. Usually, the most important source of information about a medicinal product is the Summary of product characteristics (SPC).

An SPC is a document approved by a legal authority during the medicinal product marketing authorization. It is a widely used source for retrieving information about drugs and medicinal products. An SPC has a defined structure of sections which includes sections about composition, indication, interactions, adverse effects, etc. The whole guidelines on SPCs can be found in [8].

Although an SPC is a comprehensive source of information about a medicinal product, there might be some information missing. A study [2] shows, that some new evidence about drug interactions is not included in its SPC. Similar study [16] has shown that food–drug interaction also does not have to be up to date. Despite all the imperfections, SPCs still have a great information value for physicians and are the legal basis of drug prescription in the European Union.

Unfortunately, all the information is locked in free text and no part of an SPC is delivered in a machine readable form. Combined with the length of an SPC

* This work is partially supported by the grants GAUK 572212 and GACR 204/13/08195S.

(usually about 10 pages of text) it is hard for a physician to find a particular piece of information quickly. A better format of an SPC would be useful and the idea of e-SPC was proposed in [11] where the importance of instantly available electronic drug information, which can interface with electronic health record and decision support system, is emphasized. The practice in the U.S. shows that it is possible to produce documents with drug information in a structured form – FDA Labels (structured product labeling, SPL [3]) are produced as XML documents.

In this paper we present an approach to representing and publishing SPCs as Linked Data which is a part of our bigger project called *Drug Encyclopedia*[4] which builds an information source for physicians. We take the existing SPCs, divide them into sections and annotate them with available Linked Data dictionaries. Using the semantics of each section, we are able to derive new information about, e.g., interactions. For a particular medicinal product, we can display the list of annotated entities which simplifies navigation and offers useful information easily. We are not trying to extract all information with all semantics. We rather make the search of free text easier for physicians. Then it could serve better as a part of a clinical decision support system.

Last but not least, we have built an application (based on HTML5 and RDFa) that allows to browse the annotated SPCs. Currently, we are working with SPCs in the Czech language but similar approach can be used also in different languages. Only the language modules have to be substituted.

Data published in this project can be accessed through the application at `http://datlowe.org/EncyclopediaApp` and some of it is also available through a SPARQL endpoint `http://linked.opendata.cz/sparql`.

The paper is organized as follows. In Section 2 we discuss the related work in the field of information extraction from medical texts (especially SPCs) and Linked Data in biomedical domain. Brief description of data used for annotation is delivered in Section 3. The task of SPC processing and annotation is described in Section 4. The resulting RDF representation of SPCs and extracted information is described in Section 5. The developed application for browsing Linked Data representation of SPCs is introduced in Section 6. In Section 7 we discuss the results of our research and their potential impact.

2 Related work

Structuring the content of an SPC is not a new problem. There are papers discussing structuring of particular sections e.g., pharmacodynamics section for antibiotics was examined and structured in [7]. Several papers presented approaches to extraction of drug interactions. Extraction of drug interactions based on shallow linguistics from biomedical texts in general was presented in [17]. This work concentrates on the existence of interactions and does not extract

[3] `http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/default.htm`

[4] `http://datlowe.org/drug-encyclopedia`

any other information. On the contrary, the approach from the paper [13] uses machine learning for drug interactions extraction from an SPC and also tries to get some additional information about the interaction, not only its existence. A methodology for automatic recognition of drug-related entities (active ingredient, interaction effects, etc.) based on machine learning combined with a rule based approach was presented in [14] where two levels of extraction were used.

The Linked Data principles [3] have been used for publishing data on the Web and the so called Web of Data has started to rise [4]. The biomedical data became an important part of the Web of Data. Linking Open Drug Data (LODD), which is a task force within the World Wide Web Consortium (W3C), aims on publishing drug related data as Linked Data and the participants of this task force have made twelve data sets [15] available. The biomedical Linked Data can be used in hospitals for enrichment of their own data which has been shown in a study by Mayo Clinic [12].

The Linked Data can be also used for information extraction. A prototype implementation of Lodifier converting natural language into Linked Data has been described in [1]. Extraction of medication information from discharge summaries was briefly described in [10].

DailyMed[5], source of SPLs, is included in the original LODD dataset release. Recently, a new Linked Data resource of SPLs called LinkedSPLs[6] emerged. It follows the work [5]. However, to our best knowledge, there is no other work dealing with Linked Data representation of European style SPCs.

3 Drug Linked Data Cloud

As a part of the *Drug Encyclopedia* project, an integrated data set called Drug Linked Data Cloud (DLDC) (see Figure 1) of drug related data has been created. This data set was build by linking various smaller data sets which were carefully selected on the basis of physicians requirements. Every data set contains specific information:

- **Drugs registered in the Czech Republic** - data provided by the State Institute for Drug Control (SIDC) about medicinal products marketed in the Czech Republic
- **SPC (Summary of Product Characteristics)** - documents attached to each marketed medicinal product
- **MeSH (Medical Subject Heading)** - a reference dictionary for linking other sources. It is partially translated to Czech
- **NDF-RT (National Drug File - Reference Terminology)** - data about indications, contraindications and data about pharmacological effects
- **DrugBank** - data about interactions and descriptions of active ingredients
- **ATC Hierarchy (Anatomical Therapeutic Chemical Classification System)** - classification of drugs
- **NCI Thesaurus** - direct links to FDA SPL

[5] `http://dailymed.nlm.nih.gov/dailymed/about.cfm`
[6] `http://dbmi-icode-01.dbmi.pitt.edu/linkedSPLs/`

– **FDA SPL (FDA Structured Product Labeling)** - pregnancy category
– **MedDRA (Medical Dictionary for Regulatory Activities)** - adverse event classification dictionary

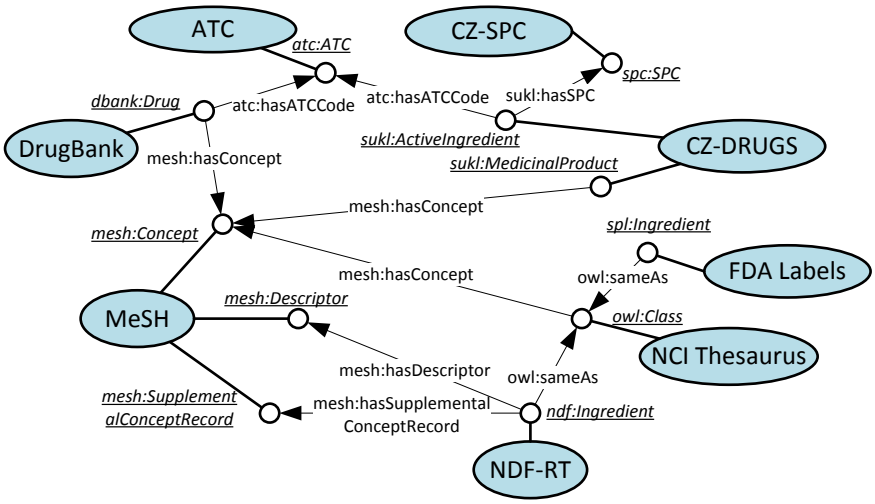

Fig. 1. Architecture of DLDC

Information contained in the briefly described data sets are valuable for physicians. But they still require the possibility of checking the information in an SPC. Therefore we have decided to use some of the structured data from DLDC for annotation of the SPCs and its structuring. We have selected Czech labels for the following entities:

– descriptors, concepts and terms from MeSH,
– active ingredients and groups of medicinal products from SIDC,
– all concepts from ATC,
– all concepts from MedDRA excluding lowest level terms.

4 SPC processing

The analysis of SPC documents covers three different tasks: section identification, annotation of text with Linked Data dictionaries, and extraction of drug interaction types. Several NLP tools and frameworks were used in our approach including the GATE framework[7], the Treex framework[8] and the Czsem tools[9].

[7] Open source software for text processing; `http://gate.ac.uk/`
[8] NLP software system; `http://ufal.mff.cuni.cz/treex/`
[9] Package integrating Treex into GATE; `http://czsem.berlios.de/`

4.1 Section identification

As already mentioned in Section 1, SPCs have a standardized structure of 28 sections. The correct identification of these sections in a document is the key for further processing and document visualization.

On the first sight, the problem looks simple. We can select lines containing the text of the SPC headlines and the text between two consecutive headlines belongs to corresponding sections. But in our case, the situation is more complicated because the documents are in Czech and various translations of the same SPC headlines from the official guideline are used in different documents. Moreover, there are spelling errors and misnumbered or omitted headlines and sections.

Our solution is based on a list of most common translations of SPC headlines taken from a collection of "training" documents. These translations are looked up in each document using an algorithm based on the well known Levenshtein edit distance so that for each SPC headline the most similar line of the document is selected. There is also a similarity threshold preventing omitted headlines to be wrongly matched with some remotely similar line.

We evaluated this approach using two different collections (training and testing) of 150 documents. Only 12 of 4198 headings (0.3 %) were missed in the testing collection and only one heading was assigned incorrectly.

4.2 Annotation of text with Linked Data dictionaries

Annotation of text with Linked Data dictionaries or *semantic annotation* is already a well established NLP task (e.g., [18]). It is conventionally solved by a gazetteer based approach[10] followed by a statistical disambiguation of ambiguous terms, see e.g., the early work of [6].

Our approach is a slight modification of the traditional one, the difference is twofold:

1. Because Czech is a flexitive language with rich morphology, term lookup without lemmatization or stemming would result in poor performance (see the comparison bellow). This problem can be elegantly solved using GATE Flexible Gazetteer [11] and a lemmatizer. Terms from gazetteer list are then matched against tokens lemmas instead of their original forms. This also implies that the gazetteer's terms have to be in the form of lemmas; therefore lemmatization was performed on each gazetteer list during its construction.
2. We are not aware of any ambiguous medical terms in gazetteer lists which we are using. The ordinary ambiguity of words (river bank / bank as institution) is solved by the lemmatizer. Therefore any disambiguation is not performed in our case.

[10] Gazetteers provide a list of known entities for a particular category, such as all countries of the world or all human diseases and are often used in information extraction. See also: `http://gate.ac.uk/userguide/chap:gazetteers`

[11] `http://gate.ac.uk/userguide/sec:gazetteers:flexgazetteer`

Lemmatization was done using the Czech morphological tagger Featurama[12], which is available through the Treex framework and integrated with GATE using Czsem tools.

We compared the results of the gazetteer lookup with and without lemmatization on the collection of 150 documents previously called "training" and found out that only 47 % of terms found using lemmatization were found also without it. It is also important to notice that lemmatization is a source of possible errors. In the case of our training collection, 2 % of terms found without lemmatization were missed by the lemmatization approach. This can be avoided by combining both approaches. Last but not least, lemmatization brings certain loss in time performance. Lemmatization took 99.5 % of the time spent on the analysis. But we can afford it since the analysis is performed offline and there is also a possibility to replace the current lemmatizer by a faster alternative (e.g. some simpler stemmer) in the future.

4.3 Extraction of drug interaction types

Task definition Although drug interactions can already be roughly inferred from the presence of semantic annotations in section *4.5 Interaction with other medicinal products and other forms of interaction* which we will call SPC interaction section (see Section 5.1 for details), we decided to provide the user with more precise information about the actual type of interaction.

Our definition of the task of extraction of interaction types is simpler than common event or relation extraction tasks known from Automatic Content Extraction[13] or similar events. We request only a valid type of interaction to be assigned to each semantic annotation present in the SPC interaction section. Current solution distinguishes only three types of interactions:

- INCREASING – simultaneous medication has amplified effect,
- DECREASING – simultaneous medication has diminishing effect and
- DENYING – the text explicitly denies such interaction.

Solution The previous two tasks – section identification and semantic annotation – were solved using shallow linguistics only. The situation is completely different in the case of this task as we used deep linguistic parsing (DLP). We have quite a rich experience with DLP and especially for Czech – a language with free word order – DLP offers a substantial simplification of rule based information extraction because it provides useful generalization of synonymous phrases and extraction techniques do not have to handle as many irregularities.

The disadvantages of DLP are low time performance and imperfection (error rate). Although time requirements of DLP are much higher than lemmatization (analysis of all SPCs took more than one week), it is still bearable in offline processing and it can be also easily speeded up by parallel processing. Error rate of DLP depends mainly on the domain from which the texts came from. SPCs

[12] `http://sourceforge.net/projects/featurama/`
[13] `http://www.itl.nist.gov/iad/mig/tests/ace/`

are written in literary language so the errors are not frequent, however numerous uncommon words (medical terms) harm the results a bit.

Our approach uses extraction rules based on the structure of (deep) syntactic trees produced by DLP from text. We use so called tectogrammatical[14] trees produced by the Treex framework. See a simplified example in Figure 2.

Note that these trees are dependency based, which means that edges connecting individual nodes are (up-down) oriented according to the grammatical importance of corresponding words.

The extraction algorithm we used can be described with the following structure and illustrated with the help of Figure 2. This algorithm applies to every semantic annotation present in the SPC interaction section. In our example, "human albumin" was annotated in the sentence.

1. Select the highest node in the corresponding tree (important for multi-word annotations, for "human albumin" it is node 4).
2. Connect this node with the closest verb up in the tree. (node 8)
3. Look if the verb and its close dependents match with some of predefined patterns. (Nodes 2 and 8 are the verb and the dependent.)
4. If matches, assign corresponding interaction type to the semantic annotation. (DENYING is attached to "human albumin" in the example case.)

Following section describes the process of discovering the predefined patterns in our solution.

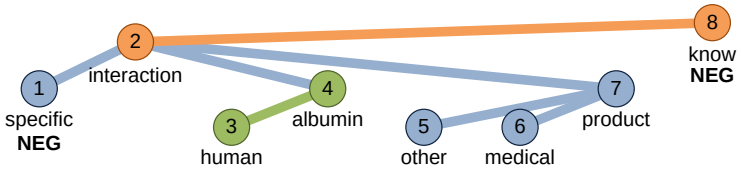

Fig. 2. Linguistic tree for sentence: *No non-specific interactions of human albumin with other medicinal products are known.*

Pattern selection The general algorithm structure described above has to be supplied with predefined verb patterns for each interaction type. These patterns were constructed manually with the help of pattern frequency analysis. The pattern frequency analysis had similar structure as extraction rules, except steps 3 and 4 were replaced with counting of occurrences of different verbs and their children nodes. The result was a sorted list of most frequent pairs "verb–child". From there it was not difficult to select the most relevant patterns for extraction rules. But it has to be noted that the list was really huge (more than 10,000 different pairs occurred at least ten times in the SPC interaction section) and we were able to go through the first 1,000 most frequent patterns only. From

[14] `http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch02.html`

there, we selected about 100 interesting patterns and generalized them to about 5 meta-rules for each interaction type. Some concrete examples can be written using following schemas. Note that the second one applies in Figure 2.

`[increase;intensify;strengthen;potentiate]+[effect;level;value; effectiveness](`*dependency types*`) -> INCREASING`

`[find;expect;observe;discover;know;demonstrate](NEG) +[interaction;influence;effect;inhibition;induction]-> DENYING`

Discussion Our approach is based on manually created extraction rules, which in contrast to machine learning approaches, has the benefit that manually annotated training collection is not needed. The evaluation of extracted interactions and interaction types is described in Section 5.2.

5 RDF representation

All data obtained during SPC processing was exported to RDF (Resource Description Framework). Respecting the Linked Data principle about reusing existing ontologies, the SALT framework (described in [9]) has been used for document and annotation representation. The SALT framework consists of three ontologies - document, annotation and rhetorical. In this work, we use document and annotation ontologies.

The document ontology describes the document structure. According to the ontology, every document may contain sections, paragraphs, sentences and text chunks. Originally, the document ontology contains only references to the original text. In our approach, we added a property `hasText` which contains a string literal with a particular part of the original text.

The annotation ontology describes the annotation structure. To every text chunk in the text an annotation can be added. Each annotation then refers to a topic represented as a URI of an annotated entity from DLDC using the property `hasTopic`.

Moreover, we have developed a model for description of other information about annotations. We are using this model to describe interaction types annotated in the SPC interaction section. An interaction type is represented as an annotation of a text chunk which was previously annotated with an entity from DLDC. This interaction type annotation also preserves the information about how it was originally derived from an SPC. Therefore, it links itself to a descriptive text chunk containing the original text which was matched by the pattern rule (see Section 4.3). The whole structure of the SPC representation including the annotations can be seen in Figure 3.

5.1 Interaction reasoning

By combining the section semantics and annotations contained in a particular section we are able to determine new relations between entities. We have used

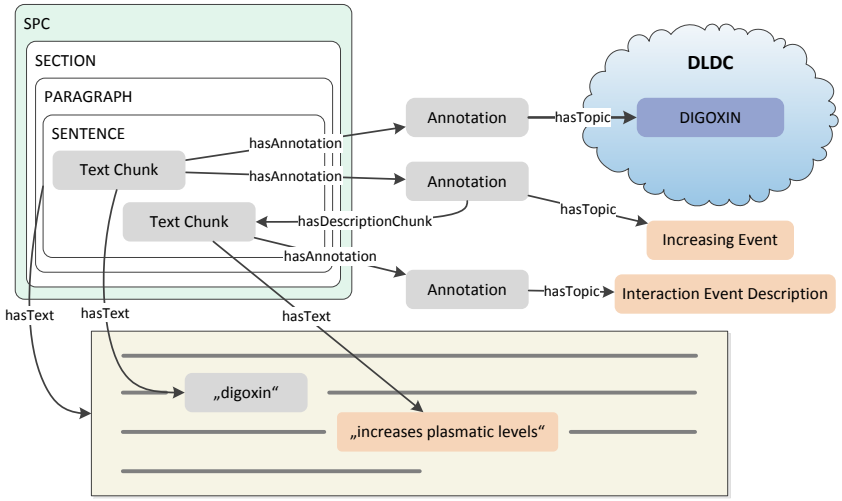

Fig. 3. SALT framework used for SPC processing.

the SPC interactions section for interaction extraction. The algorithm for the extraction is following:

> For each medicinal product containing only one active ingredient A, take each active ingredient B annotated in the SPC interactions section and create an interaction object I which refers to A and B. Use the sentence containing the annotated active ingredient B as a description of the interaction I. If the information about the interaction type is available, connect the annotation of the interaction type to the interaction I.

This rule generated many potentially interacting couples of active ingredients. Moreover, we were able to attach information about interaction type when it was available. More information follows in the next section.

5.2 Evaluation

At the time of writing of this paper, we have processed 6 764 SPCs. The resulting data set of SPCs contains 86 943 993 triples in total. It contains 187 629 extracted sections, 4 683 123 text chunks and 13 242 883 annotations.

The statistics of interaction counts can be found in the Figure 4. Using the rule for interaction extraction we have discovered 9 075 potential interactions between active ingredients. We were able to attach interaction type to 1 751 distinct interactions. Some of the interactions have at least two different interaction types attached which does not have to be wrong with respect to our rules. A relatively high number of records about nonexistence of the interaction (DENYING type) is not a problem because it is a type of information which is also helpful for physicians.

Some of the interactions are already described in other data sets from DLDC. DrugBank and discovered interactions have 706 of them in common. NDF-RT and discovered interactions have 533 of them in common. And at last, all three mentioned data sources have 381 interactions in common.

Regarding the SPC processing we believe we will be able to refine the process and discover more relevant information about interactions in future work.

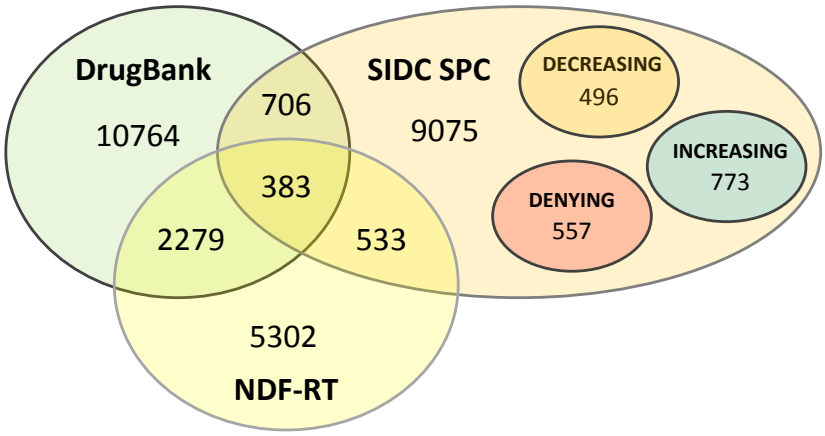

Fig. 4. Statistics of interaction counts in data sets in DLDC and extracted interactions from SPCs

6 Usage Scenarios

There are many ways how Linked Data representation of SPCs can be exploited at the application layer and bring benefits to end users. In this section, we demonstrate two usage scenarios. The first scenario shows how the SPCs enriched with data about related medicinal products, drug ingredients, diseases, etc. from DLDC can be browsed by users. A user can search for an SPC of a chosen medicinal product, see the entities annotated in the text and display the data about those entities. From a displayed entity, the user can browse to other related entities and through those he can access other related SPCs. In other words, the user can freely move between SPCs and related entities.

The second scenario is based on the new data about the entities annotated in SPCs extracted from the textual content of SPCs (see Section 4.3). The new data can be e.g., new interactions and side effects of drug ingredients. We show that it is necessary to keep mappings between the new data and the textual content from where the data was extracted. It enables a user to easily see this particular part of the text (e.g., the sentence from where a new interaction was extracted) and to verify the extracted data. Moreover, thanks to the mappings, the newly

extracted data enriched with the other data in DLDC serves as an index. This index structure is much more "intelligent" than a full-text index. For example, using a full-text index, it is not possible to search for all interactions between two particular groups of ingredients when only concrete ingredients are present in the text. Our index "knows" whether those ingredients belong to the requested groups and what particular parts of the text (e.g., particular sentences) describe the requested interactions. Therefore, it is able to answer such questions.

We have implemented both scenarios in our experimental application *Drug Encyclopedia*.

6.1 Browsing Enriched SPCs

The basic requirement is that it must be possible to display an SPC of a chosen medicinal product. Our RDF representation also contains information that a particular part of the text, the text chunk, has a meaning specified by annotations attached to the text chunk. Therefore, it is possible to display the SPC with highlighted annotated text chunks. When a user reads the SPC he or she can easily see the highlighted text chunks. A text chunk can be annotated with more meanings. A meaning means mapping of the text chunk to a particular entity in DLDC. The user can then display those meanings. If the user is interested in the detailed data about a particular meaning, (s)he can display it.

To implement such a functionality in the *Drug Encyclopedia* application we need to represent SPCs in a format which enables to keep the original text of the SPC and add the annotations to the original text. These annotations are represented in RDF format as described in Section 5. We show that HTML5+RDFa format is suitable for such representation. HTML5 provides a special semantic tag `article` which enables to represent a textual content of an article - in our case a textual content of an SPC. The element is further structured using `section` and `paragraph` elements. These elements enable us to represent the basic identified structure of an SPC with standardized HTML5 constructs. Each identified text chunk is then denoted with a `span` element. A sample HTML5 document with a part of an SPC document is displayed in Figure 5.

With RDFa attributes which extend HTML5 elements, we further enrich the SPC encoded in HTML5 with its RDF representation – here each structural part of the SPC (i.e., sections, paragraphs and text chunks) has its own URI which is an identifier and a mean of access to the part. The URI is encoded as an RDFa attribute `resource` of the respective HTML5 element. We also use `typeof` RDFa attribute to encode the type of the resource. If the resource is an object of some RDF triple with the resource represented by the parent element as a subject, we encode this property with `property` RDFa attribute. Examples of these RDFa attributes can be seen in Figure 5.

Each text chunk identified in the SPC has one or more annotations which associate the text chunk with respective entities in DLDC. These annotations are not part of the textual content of the SPC. Therefore, they can not be encoded directly in the text. HTML5 offers a special element `footer` where additional data for an article can be provided. We use this element to encode annotations

```
<article> ...
  <section property="sdo:hasSection" typeof="sdo:Section"
      resource="spc/section/SPC98910-4">
    <h1 property="sdo:hasSectionTitle">4. Clinical information</h1> ...
    <section property="sdo:hasSubSection" typeof="sdo:Section"
        resource="spc/section/SPC98910-4-5">
      <h1 property="sdo:hasSectionTitle">4.5. Interactions</h1> ...
      <p property="sdo:hasParagraph" typeof="sdo:Paragraph"
          resource="spc/paragraph/SPC98910-4-1">
        ... <span property="sdo:hasTextChunk" typeof="sdo:TextChunk"
                resource="sukl/spc/text-chunk/SPC98910-4-5-1-4-8720-8728">
              digoxin</span> ... </p>
    ... </section> ... </section> ...
  <footer>
    ...
    <div about="sukl/spc/text-chunk/SPC98910_doc-4-5-1-4-8720-8728">
      ...
      <div property="sdo:hasAnnotation" typeof="sao:Annotation"
          resource="sukl/spc/annotation/SPC98910_doc-4-5-1-4-8720-8728-
                    M0006386-enc">
        <div property="sao:hasTopic" typeof="enc:Ingredient"
            resource="drug-encyclopedia/ingredient/M0006386">
          <span property="dcterms:title">digoxin</span>
        </div></div> ... </div> ... </footer></article>
```

Fig. 5. Sample RDFa representation of an SPC

of text chunks. For each text chunk there is an HTML element `div` with the URI of the text chunk as a value of its RDFa attribute `about`. This attribute means that data nested in this `div` element further extends data about the text chunk. We therefore nest the annotations of the text chunk here. Each annotation is encoded as a nested `div` element. It further contains another `div` element which encodes the topic of the annotation, i.e. the reference to the related entity in DLDC. Our sample displayed in Figure 5 demonstrates RDFa representation of annotations of the text chunk *digoxin*.

The *Drug Encyclopedia* web application contains a component which is able to display SPCs encoded in HTML5+RDFa format as described above. It highlights text chunks and when a user moves a mouse pointer over a highlighted text chunk it shows a box with the annotations. Each annotation is displayed as an active link to the detail of the respective entity from DLDC. This enables the user to browse from the SPC to related entities and from here to other related SPCs. Sample screen shots of the application, which demonstrate how we implemented such browsing, are displayed in Figure 6.

6.2 Searching in Semantically Indexed SPCs

Displaying annotations of text chunks and enabling navigation to the detailed data about the annotated entities is only a basic usage scenario. As we have described in Section 4.3, it is also possible to extract new data from the textual content of SPCs. We can extract drug interactions, side effects or therapeutic indication for a medicinal product or drug ingredient. All this data further enriches the data about entities we already have from other structured data sources in

Fig. 6. Screenshots from *Drug Encyclopedia* application: (1) page displaying an SPC encoded in HTML5+RDFa and a box with annotations of a text chunk *Digoxin* chosen by a user and (2) another page displaying a detail of *Digoxin* ingredient after a user clicks on one of the annotations of *Digoxin* text chunk.

DLDC. We are then able to create a record for each entity which contains related new data extracted from SPCs and provide links to the original textual content of SPCs where the user can read detailed text which explains the extracted data in a more detail. Since the process of extraction of the new data can not be correct in all cases (automatic methods always contain some errors), such records should be used jointly with the original textual content of SPCs. For example, a user can search for SPCs and their parts which explain interactions between two drug ingredients or ingredient groups. In other words, records are useful for users since it enables them to more easily and quickly find required information. Section 5 shows how relationship between derived interactions and the original text is recorded in RDF representation. Figure 7 shows how we display these relationships to users of *Drug Encyclopedia*. There is a screen shot which displays an interaction of *Ibuprofen* with *Phenobarbital*. There is a description of the interaction highlighted. This description is a sentence from an SPC document where the interaction has been identified. Our application enables the user to display the whole section describing interactions of *Ibuprofen* where he or she can read more detailed information.

7 Conclusions

In this paper we have presented a possible approach for representation of SPCs. It could serve as an example for authorities how users (typically physicians) can benefit from better structured SPCs. It offers faster navigation through sections. And the combination with annotations which are a higher form of structuring the information contained in SPCs, the SPCs can be used together with structured information from other data sources.

Fig. 7. Screenshots from *Drug Encyclopedia* application: (1) page displaying an interaction between two ingredients *Ibuprofen* and *Phenobarbital* extracted from textual content of an SPC and (2) page displayed after the user requests the original textual content of an SPC from where the interaction was extracted.

The whole project of *Drug Encyclopedia* is intensively consulted with physicians. It was the same with the processing of the SPCs we have presented here. The outputs are welcomed by the physicians we communicate with. We plan to distribute our application among the whole community of physicians, get feedback and proper evaluation of user experience.

In future we will work on the improvement of the SPC processing in order to get more accurate annotations and therefore more precise information about, e.g. interactions. There is also a possibility to dive into other sections, e.g. indication to extract more structured information which could be helpful for physicians. At last but not least, we are planning to process patient information leaflets and therefore offer the application also to patients.

References

1. I. Augenstein, S. Padó, and S. Rudolph. Lodifier: Generating linked data from unstructured text. In *The Semantic Web: Research and Applications*, volume 7295 of *Lecture Notes in Computer Science*, pages 210–224. Springer Berlin Heidelberg, 2012.
2. V. Bergk, W. Haefeli, C. Gasse, H. Brenner, and M. Martin-Facklam. Information deficits in the summary of product characteristics preclude an optimal management of drug interactions: a comparison with evidence from the literature. *European Journal of Clinical Pharmacology*, 61(5-6):327–335, 2005.
3. T. Berners-Lee. Linked Data. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006. Accessed: 2013-09-16.
4. C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

5. R. Boyce, J. Horn, O. Hassanzadeh, A. d. Waard, J. Schneider, J. Luciano, M. Rastegar-Mojarad, and M. Liakata. Dynamic enhancement of drug product labels to support drug safety, efficacy, and effectiveness. *Journal of Biomedical Semantics*, 4(1):5, 2013.
6. S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien. SemTag and seeker: bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 178–186, New York, NY, USA, 2003. ACM.
7. C. Duclos, G. L. Cartolano, M. Ghez, and A. Venot. Model Formulation: Structured Representation of the Pharmacodynamics Section of the Summary of Product Characteristics for Antibiotics: Application for Automated Extraction and Visualization of Their Antimicrobial Activity Spectra. *JAMIA*, 11(4):285–293, 2004.
8. European Medicines Agency. A guideline on summary of product characteristics. `http://ec.europa.eu/health/files/eudralex/vol-2/c/smpc_guideline_rev2_en.pdf`, 2009. Accessed: 2013-09-16.
9. T. Groza, S. Handschuh, K. Möller, and S. Decker. SALT - Semantically Annotated LaTeX for Scientific Publications. In E. Franconi, M. Kifer, and W. May, editors, *The Semantic Web: Research and Applications*, volume 4519 of *Lecture Notes in Computer Science*, pages 518–532. Springer Berlin Heidelberg, 2007.
10. J. Kozák, M. Nečaský, and J. Pokorný. Extracting Medical Information Using Linked Data. In *SWAT4LS*, volume 952 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.
11. S. Maxwell, H.-G. Eichler, A. Bucsics, W. E. Haefeli, L. L. Gustafsson, and the e SPC Consortium. e-SPC – delivering drug information in the 21st century: developing new approaches to deliver drug information to prescribers. *British Journal of Clinical Pharmacology*, 73(1):12–15, 2012.
12. J. Pathak, R. C. Kiefer, and C. G. Chute. Applying linked data principles to represent patient's electronic health records at mayo clinic: a case report. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, pages 455–464, New York, NY, USA, 2012. ACM.
13. S. Rubrichi and S. Quaglini. Summary of product characteristics content extraction for a safe drugs usage. *Journal of Biomedical Informatics*, 45(2):231 – 239, 2012.
14. S. Rubrichi, S. Quaglini, A. Spengler, P. Russo, and P. Gallinari. A system for the extraction and representation of summary of product characteristics content. *Artificial Intelligence in Medicine*, 57(2):145 – 154, 2013.
15. M. Samwald, A. Jentzsch, C. Bouton, C. Kallesœ, E. Willighagen, J. Hajagos, M. Marshall, E. Prud'hommeaux, O. Hassenzadeh, E. Pichler, and S. Stephens. Linked open drug data for pharmaceutical research and development. *Journal of Cheminformatics*, 3(1):1–6, 2011.
16. M. T. San Miguel, J. A. Martínez, and E. Vargas. Food–drug interactions in the summary of product characteristics of proprietary medicinal products. *European Journal of Clinical Pharmacology*, 61(2):77–83, 2005.
17. I. Segura-Bedmar, P. Martínez, and C. De Pablo-Sánchez. Using a shallow linguistic kernel for drug-drug interaction extraction. *J. of Biomedical Informatics*, 44(5):789–804, Oct. 2011.
18. C. Tao, D. Song, D. Sharma, and C. G. Chute. Semantator: Semantic annotator for converting biomedical text to linked data. *Journal of Biomedical Informatics*, 46(5):882 – 893, 2013.