Drugs and Disorders:
From specialized resources to Web data

Caroline Barrière[1], Michel Gagnon[2]

[1] Centre de Recherche Informatique de Montréal, 405 Ogilvy suite 101,
Montréal, QC, Canada, H1M 1N3
[2] École Polytechnique de Montréal,
C.P. 6079, succ. Centre-ville
Montréal, QC, Canada, H3C 3A7

caroline.barriere@crim.ca, mgagnon@polymtl.ca

Abstract. In this article, we focus on the *may_treat* predicate linking drugs and disorders. Such predicate is expressed in RDF format in the VHA National Drug File Reference Terminology (NDF-RT), a specialized medical resource. The DailyMed dataset also contains this predicate, but only in textual form: for each drug there is an *indication* field that links the drug's URI to a literal that is a long description (more than 100 words on the average). We show that natural language processing (NLP) techniques can be used to further distil the indication field to extract *may_treat* predicates. We then move to Web exploration and show how we can apply similar NLP techniques to find *may_treat* predicates. The diversity in natural language expressions and the embedding of good information among noisy and often redundant data make it a challenge to exploit the Web. Still, we show that it can be used for finding NDF-RT *may_treat* predicates with comparable success to using DailyMed.

Keywords: linguistic patterns, NDF-RT, UMLS, DailyMed, text mining, Web as corpus.

1 Introduction

The number of sources for medical data on the Web is large and growing every day. One only has to look at the BioPortal site [1] to see the number of available medical ontologies, or to type in a disease name in Google to find multiple sites about this condition. Many resources exist in non-RDF forms as public databases that can be downloaded (UMLS, Snomed, Mesh)[1] and others exist in RDF. In the Linked Data Cloud[2], information about drugs and diseases can be found in many datasets such as: DbPedia[3], Drugbank, DailyMed, Diseasome, Medicare, SIDER[4].

[1] These three resources are available at http://www.nlm.nih.gov/research/ .
[2] http://linkeddata.org/
[3] http://dbpedia.org/
[4] http://www4.wiwiss.fu-berlin.de/ provides access to Drugbank, DailyMed, Diseasome, Medicare and SIDER.

In this research, we explore data sets of different levels of specialization and different levels of RDFization to reflect on possible ties with research in Corpus Linguistics. To focus this study on the distinction between datasets, we look at one single predicate, the *may_treat* predicate linking drugs and disorders.

In section 2, we introduce the National Drug File Reference Terminology (NDF-RT) and the Unified Medical Language System (UMLS). Throughout this study we rely on *may_treat* predicates from NDF-RT, and on UMLS concept identifiers and labels for subjects and objects of the *may_treat* predicates.

In section 3, we look at the DailyMed dataset, and how the *may_treat* predicate is rather expressed in textual form. For each drug, DailyMed contains an *indication* field[5], a literal that is a long description (more than 100 words on the average). We can apply natural language processing techniques to further analyze this textual information and generate one or more *may_treat* predicates from it.

In section 4, we then contrast such resource with the Web at large, and show how the diversity in natural language expressions and the embedding of purposeful information among noisy and often redundant data makes it a challenge to exploit. Still, we argue that Web data is valuable, and can help expand specialized resources such as the NDF-RT.

In section 5, we conclude on this exploration by comparing the two resources and point to many possible ways to expand our search, either in DailyMed or on the Web at large, but taking different strategies based on the difference in the resources.

2 NDF-RT and UMLS

Not part of the Linked Data Cloud, the National Drug File Reference Terminology (NDF-RT)[2][3] describes and defines medications. More specifically it describes generic ingredients or combinations, providing their active ingredients, mechanisms of action, physiologic effects, indications and contraindications.

NDF-RT has a distribution in OWL[6]. It contains more than 44000 concepts, each one with a link to a UMLS concept unique identifier (CUI). Not part of the semantic web (although referred to in many sites), UMLS contains over 2 million names for some 900 000 concepts from more than 60 families of biomedical vocabularies, and 12 million relations among these concepts [4]. Each concept in UMLS has a unique id (CUI), and is associated to a set of labels. These labels are essential to perform text analysis, as they give many possible ways of matching concepts by their lexical expression.

This greatly enriches the NDF-RT by providing many labels for each of its concepts. We calculated that on average, drugs have 6 labels and diseases have 19 labels. Table 1a and 1b show examples of drugs and disorders with various labels. Their corresponding rows form *may_treat* pairs.

[5] To be correct, it should be called a predicate, as DailyMed is available in RDF format, but the term “field” is used throughout this article to differentiate between the *indication* long literal (textual data) of DailyMed and the *may_treat* predicate in NDF-RT.

[6] The US government provides quarterly updates of the terminology in a variety of formats (XML, OWL, and text) at http://evs.nci.nih.gov/ftp1/NDF-RT/.

Table 1a. Drugs with multiple labels found from UMLS

CUI	Examples of Labels
C0980568	Theophylline, anhydrous 200mg capsule THEOPHYLLINE 200 MG ORAL CAPSULE, EXTENDED RELEASE THEOPHYLLINE ANHYDROUS 200 MG ORAL CAPSULE, EXTENDED RELEASE Theophylline, anhydrous 200mg capsule (product)
C0980014	RISPERIDONE 3 MG ORAL TABLET Risperidone 3mg tablet Risperidone 3mg tablet (product) Risperidone 3mg tablet (substance) RISPERIDONE 3 MG ORAL TABLET, FILM COATED RISPERIDONE 3 MG ORAL TABLET, ORALLY DISINTEGRATING

Table 1b. Disorders with multiple labels from UMLS

CUI	Examples of Labels
C0003578	Apnea / APNEA / Apneas / Apnoea / RESPIRATORY ARREST / ARREST, RESPIRATORY / Apnea / Apnoea / Has stopped breathing / Not breathing / Apneic / Apnoeic
C0040517	Gilles de la Tourette syndrome / Gilles de la Tourette's syndrome Tourette's disorder / Tourette Disorder / Syndrome, Tourette's Tourette's Syndrome / Tourettes Syndrome / Tourette's Disease Tourette Disease / Tourettes Disease Combined Multiple Motor and Vocal Tic Disorder Combined Vocal and Multiple Motor Tic Disorder

UMLS also contains a Semantic Network, which defines 54 relationships as well as 133 semantic types organized in 11 semantic groups. Two specific semantic groups interest us: "Chemical & Drugs" (which we often refer to as "drugs" in this article) and "Disorders" as they respectively represent the subject and object domains for the *may_treat* predicate. Table 2 shows statistics about the NDF/RT concepts with their corresponding semantic groups defined in UMLS. The distribution certainly reflects the focus of NDF/RT on drugs and disorders.

Table 2. Distribution of some UMLS Semantic Groups in NDF/RT

Semantic Group	Nb Concepts
Activities & Behaviors	1
Anatomy	27
Chemical & Drugs	31527
Concepts & Ideas	41
Devices	649
Disorders	9115
Genes & Molecular Sequences	16
Living Beings	48
Objects	622
Phenomena	10
Physiology	2342
TOTAL	44398

Among the 31527 concepts in NDF-RT associated with the semantic group "Chemical & Drugs", we find that only 8836 (28%) of these concepts participate in the *may_treat* relation. And among the 9115 concepts in NDF-RT associated with the semantic group "Disorders", 962 (11%) participate in the *may_treat* relation.

These small percentages show how such specialised resource is very valuable, but also limited in its coverage. All drugs could be involved in a *may_treat* relation, but only 28% of them actually are at this time (for this version of the resource[7]).

3 DailyMed

The site DailyMed (http://dailymed.nlm.nih.gov) published by the National Library of Medicine provides high quality information about market drugs. A Linked Data version provides a RDF view of part of the information at http://www4.wiwiss.fu-berlin.de/dailymed/. The RDF version of DailyMed is part of the Linking Open Drug Data project [5]. It describes about 3600 drugs and provides many predicates, among which, some lead to resources and others to literals.

Some predicates in the RDF view link to resources, and others to literals of variable sizes. Predicates such as "adverseReaction", "clinicalPharmacology", "precaution" or "indication" lead to literals that are actually textual data on which text analysis techniques can be used to further pursue the RDFization. The indication for each drug is rather lengthy (size varies from 1 word to 1338 words, with a means of 127). Figure 1 shows some examples. In bold are linguistic patterns, as we will refer to them in section 3.3.

http://www4.wiwiss.fu-berlin.de/dailymed/resource/drugs/4292
Fluticasone propionate ointment **is** a medium potency corticosteroid **indicated for the relief of** the inflammatory and pruritic manifestations of corticosteroid-responsive dermatoses in adult patients.
http://www4.wiwiss.fu-berlin.de/dailymed/resource/drugs/4293
RYZOLT **is indicated for the management of** moderate to moderately severe chronic pain in adults who require around-the-clock treatment of their pain for an extended period of time.
http://www4.wiwiss.fu-berlin.de/dailymed/resource/drugs/4304
PhosLo **is indicated for the control of** hyperphosphatemia in end stage renal failure and does not promote aluminum absorption.
http://www4.wiwiss.fu-berlin.de/dailymed/resource/drugs/4308
Astelin Nasal Spray **is indicated for the treatment of the symptoms of** seasonal allergic rhinitis **such as** rhinorrhea, sneezing, and nasal pruritus in adults and children 5 years and older, **and for the treatment of the symptoms of** vasomotor rhinitis, **such as** rhinorrhea, nasal congestion and post nasal drip in adults and children 12 years and older.

Fig. 1. Examples from DailyMed

[7] Version NDFRT_Public_2011.07.05_TDE.xml.

3.1 Coverage of DailyMed drug names in NDF-RT

Our first challenge is to establish a correspondence between DailyMed concepts and NDF-RT concepts. Dailymed does not provide a UMLS CUI, nor any other CUIs contained in NDF-RT (such as Mesh_CUI) that could have been use as an intermediate to link to a UMLS_CUI. Matches must therefore be established via concept labels and the process becomes prone to uncertainty and errors.

We rely on Lucene[8], an open-source document indexing and retrieval software. All UMLS CUIs with their associated labels (as presented in section 2) are indexed in Lucene. All DailyMed drug names, given by the dailymed *name* property are used in turn as query.

Different retrieval strategies are implemented in Lucene and can be parameterized, but we simply use the default TF-IDF (Term Frequency-Inverse Document Frequency) scoring which considers all labels as a bag-of-words[9]. Using the matching process on all drugs, we established that of its 2305 drugs[10], Dailymed has 987 of them that are part of NDF-RT. Some examples of matching labels are shown in Table 3.

Table 3. Matching UMLS_CUIs and DailyMed drug names

DailyMed name	Found UMLS_CUI	Associated label to UMLS_CUI
Mefloquine HCl	C0025153	Mefloquine
Cardene	C0591232	Cardene
	C1240694	Cardene Capsules
Ringer's	C0073384	Ringer's acetate
	C0073385	Ringer's Solution
Captopril and Hydrochlorothiazide	C2930440	captopril, hydrochlorothiazide drug combination
	C0688113	Captopril+hydrochlorothiazide 25mg/15mg tablet

3.2 Finding *may_treat* predicates in the "indication" field

The Stanford Parser [7] is used to process indications by performing sentence splitting, tokenization, part-of-speech tagging and parsing. Using this parsing process, we are able to process 3231 indications (92.6% of all indications[11]), and from these 2090 different drugs are covered (90% of the list, from which already 3% had no indication fields).

We are currently developing approaches which will take advantage of the parse tree, as promoted in the literature, especially in BioNLP [8][9], but for the research reported here, a simpler approach is used to discover *may_treat* predicates in

[8] Lucene is available for download at http://lucene.apache.org/java/docs/index.html.

[9] Information Retrieval strategies are beyond the scope of this article, and we refer the reader to the introductory book by Manning et al. 2008 [6].

[10] The SPARQL endpoint http://www4.wiwiss.fu-berlin.de/dailymed/sparql was queried in July 2011 to obtain drugs names and indications.

[11] Some sentences were very long with long enumerations of disorders and side-effects. These, among others, represent types of sentences that the parser could not digest.

indication fields. In fact, the underlying idea is very simple and consists in finding noun phrases (NPs) corresponding to drugs and disorders. These become candidates for *may_treat* predicates to be validated against NDF-RT predicates. Taking all pairs of NPs is a bit naïve if precision is our goal, but here, recall will be measured and used as a comparison point for future Web analysis.

The Stanford Parser gives lists of NPs for each sentence. Each one can then be matched to concept labels in UMLS using Lucene. Perfect matches will be put first and then partial matching after.

As an example shown in Table 4, the first sentence of Figure 1 is processed and its NPs are matched to UMLS CUIs. The list of CUIs found from UMLS is then restricted to the ones in semantic groups "Chemical & Drugs" and "Disorders" (see the semantic group in last column of Table 4).

With this process, we find 849 drugs from DailyMed (41% of the 2090 analyzed) which participate in a *may_treat* relation in NDF-RT. From these, only 457 drugs (54%) contain a NP that can be linked to a disorder listed in *may_treat* pairs from NDF-RT for that drug.

Table 4. Noun Phrases extracted for the sentence with their closest CUIs from UMLS

	NP	CUI	Closest Labels	Semantic Type	Semantic Group
1	Fluticasone propionate ointment	C0590979	Fluticasone propionate 0.005% ointment	Clinical Drug	Chemicals & Drugs
2	a medium potency	C0009458	Communications Media	Intellectual Product	Concepts & Ideas
		C0439536	Medium	Quantitative Concept	Concepts and Ideas
		C0486821	Gentamicin.high potency	Organic Chemical	Chemical & Drugs
3	the relief	C0564405	Relief	Finding	Disorders
		C0035038	Relief Work	Occupational Activity	Activities & Behaviors
		C0451615	Pain relief	Therapeutic or Preventive Procedure	Procedures
4	the inflammatory and pruritic manifestations	C0033771	Prurigo	Disease or Syndrome	Disorders
		C0033774	Pruritus	Sign or Symptom	Disorders
5	corticosteroid-responsive dermatoses	C0037274	Dermatoses	Disease or Syndrome	Disorders
		C0015456	Facial Dermatoses	Disease or Syndrome	Disorders
		C0018567	Hand Dermatoses	Disease or Syndrome	Disorders
6	adult patients	C0030705	Patients	Patient or Disabled Group	Living Beings
		C0001675	Adult	Age Group	Living Beings

In Table 5 shows examples of pairs for which the drug is found in the *may_treat* pairs, but the disorder NPs are not found. Even without being a medical expert, these pairs seem "reasonable" and could be suggested as candidates (see last column of Table 5 for comparison). In our opinion, if the process is to be used for knowledge discovery, it is best to imagine a system in which automatic discovery is not a final step, but a step within a semi-automatic process leading to *may_treat* predicate candidates to be validated by a medical expert.

Table 5. Examples of pairs extracted from indication fields for which Drug is the subject of *may_treat* but Disorder is not its object

DailyMed drug name (NP)	drug CUI found	DailyMed disorder name (NP)	disorder CUI found	CUI label	Actual labels found in *may_treat* of NDF-RT
Sodium Chloride	C0037494	salt syndrome	C0866191	Salt-losing nephritis or syndrome	Shock, Hemorrhagic Dry Eye Syndromes Corneal edema Wounds and Injuries Hyponatremia Dehydration
Gabapentin	C0060926	neuralgia	C0027796	Neuralgia	Bipolar Disorder Epilepsies, Partial Phobic anxiety disorder Pain
Amiodarone Hydrochloride	C0700442	fibrillation	C0232197	Fibrillation	Ventricular Fibrillation Supraventricular tachycardia

3.3 Analyzing how *may_treat* is expressed

We focus now on *may_treat* pairs which were found in the indication fields to see how this information is expressed in text. There are 954 sentences in which *may_treat* pairs are found (covering 457 drugs, which we mentioned earlier). In these sentences, we simply record what occurs in-between the pairs as a possible linguistic pattern. Such possible patterns were emphasized in bold in Figure 1.

The use of linguistic patterns for knowledge discovery has been the subject of much research in corpus linguistic and terminology [10]. The general idea is to find sentences containing known relations to discover how these relations are expressed in natural language. Language is ambiguous and varied, but when specific relations are expressed, some more or less regular patterns can be discovered. Once these patterns are discovered, they can be used (with care as they are often noisy) to discover instances of relations.

All linguistic patterns recorded become candidate patterns. The weight of each one is calculated. To do so, we take into account that each indication sentence could lead to multiple *may_treat* candidate pairs, and therefore to multiple pattern candidates. A weight of 1/nbCandidates is assigned to each pattern candidate for that sentence. The total weight on all sentences for all patterns is then calculated[12]. We show the top 20 patterns in the Table 6 below.

Table 6. Patterns found between may_treat pairs in DailyMed indication fields

Pattern found	Weight
is indicated for the treatment of	40.75
are indicated for the treatment of	39.89
are indicated in the management of	21.17
is indicated in the management of	14.50
is not indicated for the treatment of	12.67
are indicated in the treatment of	11.75
is indicated in the treatment of	9.50
are indicated for the management of	9.50
is indicated for	8.13
is indicated for the control of	5.85
) is indicated for the treatment of	5.58
are indicated for the long-term management of	5.0
is indicated for the management of	5.0
is indicated for the topical treatment of	4.62
is indicated for the temporary relief of	4.61
is indicated for the relief of	3.83
is indicated for the local treatment of	3.67
is indicated for use in the treatment of	3.5
, is indicated for the treatment of	3.33
are indicated for	3.09

3.4 Conclusions on DailyMed

The exploration of Dailymed and other RDF resources to fully exploit their textual data and transform them into RDF triplets is a research topic in itself deserving more research efforts. But as the focus of this article is the exploitation of Web data, we stop here our exploration to move to the Web. We will use this exploration as a comparison point, and also as a first entry point into the noisy web data. In our brief exploration of DailyMed, we have shown that:

1. Coverage is different than NDF-RT, with only 987 of its 2305 drugs present in NDF-RT.
2. Indication fields can be analyzed with the StanfordParser to do part-of-speech tagging and retrieve NPs which can be matched to drugs and disorders.
3. Only a portion of the indications lead to known *may_treat* pairs from NDF-RT. This comparison establishes recall, but does not inform us about the value of new knowledge found. If the method is able to recall known information, we infer

[12] The size of candidate patterns is empirically set to a maximum of 50 characters before we calculate the weights. With our naïve method of finding all NPs, we generate very long patterns and need to set a size limit.

that it will also be able to find new information that can become candidate information to be reviewed by medical experts to be added in a specialized resource.

4. Label matching is not obvious and could be a research topic in itself. At the present, we rely on default TF-IDF strategies implemented in Lucene.
5. Ways of expressing the *may_treat* relation are varied but still limited and almost all patterns contain the keyword "indicated" in them. Our purpose here is to find what seems to be the most common way of expressing the *may_treat* relation. We pursue in another research project a full use of these patterns (expressed syntactically) for the purpose of precisely extracting all information within DailyMed indications.

4 Web Data

We first discuss the presence of drug information on the Web. We then look into how to find *may_treat* pairs using the single word "indicated", word common to all patterns in DailyMed. We show some positive results, as some pairs from NDF-RT can be found in such way, providing support for the method. We then talk about the common linguistic patterns for *may_treat* pairs on the noisy web.

4.1 Are drugs mentioned on the Web?

A first investigation is the actual presence of drugs on the Web. Before we even look at whether *may_treat* pairs are present, we first look at the presence of drugs themselves. Contrarily to Dailymed, we do not have a list of drugs and indications, so we must search for them using the known labels for these drugs.

As we have seen earlier, we have about 6 labels per drug and 19 labels per disease given by UMLS. We establish "presence" by finding hit counts for the labels. To find hit counts, we work with the Bing API[13]. Tables 7a and 7b show examples of hit counts for a few labels in three drugs and disorders.

We randomly chose 8000 pairs among the 47218 *may_treat* pairs in NDF-RT, and we calculated statistics on the presence of the drugs and diseases on the web. The number of different drugs in these pairs happens to be 3560, and the number of different disorders is 779. We found that all disorders have at least one label that has a presence on the web. That is not the case for the drugs, as we calculated that 72% of them (2547/3560) that have no presence on the web.

It is probably incorrect to say that the drug has no web presence at all, but by using its different labels, as found in UMLS, we are not able to access them. As future work, we can investigate looking at active ingredients or other information about them to find them.

[13] The Bing API allows web searches to be embedded in a Java program. Information and download can be found at: http://msdn.microsoft.com/en-us/library/dd251056.aspx

Table 7a. Examples of hit counts for drug labels

CUI	Label	Hit Count
C0006681	Calcium Carbonate	2380000
	CALCIUM CARBONATE	2380000
	Carbonate, Calcium	-1
	Calcium carbonate (substance)	-1
C0976710	Erythromycin 1.5 solution	-1
	Erythromycin 15mgL solution	-1
	Erythromycin 1.5 solution (product)	-1
C0724633	metoprolol succinate	255000
	Metoprolol Succinate	255000
	metoprolol CR-XL	-1
	Metoprolol succinate (substance)	-1

Table 7b. Examples of hit counts for disease labels

CUI	Label	Hit Count
C0002963	Angina Pectoris, Variant	-1
	Variant angina pectoris	113000
	Prinzmetal Angina	70200
	Coronary artery spasm angina	706000
C0026918	Mycobacterium Infections	577000
	Infection, Mycobacterium	-1
	Mycobacteriosis	41100
	Infection due to mycobacterium species	310000
	INFECT MYCOBACT	5810
C0017168	Gastroesophageal Reflux Disease	1210000
	Acid Reflux	14400000
	Gastrooesophageal reflux disease	4320
	Esophageal reflux	852000
	Oesophageal reflux	178000
	Esophageal reflux NOS	-1

Table 8 shows the distribution on number of hit counts based on frequency of different labels for drugs and diseases.

Table 8. Result of hit counts for drugs and diseases.

Hit Count	% drugs	% diseases
0	80,96%	49,90%
100	0,82%	0,56%
1000	1,35%	1,61%
10000	3,23%	5,68%
100000	5,77%	10,81%
1000000	5,42%	15,87%
10000000	1,97%	11,99%
100000000	0,47%	3,31%
1000000000	0,01%	0,27%

4.2 Building a data set for experimentation

As mentioned earlier, 8000 *may_treat* pairs were randomly selected to establish the presence of drugs and disorders on the Web. The most problematic category is the drug as they tend to be mentioned in very specific ways and a large proportion is not found on the Web.

To generate pairs for *may_treat* web exploration, a minimum hit count for either drug or disease label was set at 100000, and from those, pairs with a joint hit count of more than 10000 were selected.

Table 9 shows examples of pairs with their joint hit counts. Table 10 shows distribution of hit counts for pairs with drugs & disorders having hit counts more than 100K.

Table 9. Example of joint hit counts for random pairs selected.

Drug Label	Disorder Label	Joint hit count
KETOCONAZOLE	Cutaneous Candidiasis	649000
Metronidazole 1 gel	Giardiasis	4
Erythromycin 2 topical gel	Gonorrhea	0
danazol	endometriosis	81200
Ciprofloxacin	The clap	35600
Recombinant Interferon alpha 2b	Melanoma	106000
Hydrocortisone Valerate	Facial Dermatosis	0
Fentanyl	Unspecified pain	3170
Lidocaine 5 ointment	Drug Toxicities	0
Anastrozole	Breast Tumors	14800
Magnesium Hydrate	Dyspepsia	318
Ethinyl estradiol	Tumour of prostate	0

Table 10. Statistics on joint hit counts for pairs with individual hit counts above 100000.

Joint hit count	Nb Pairs
0	361
100	131
1000	92
10000	100
100000	175
1000000	115
10000000	25
100000000	0

4.3 Looking for *may_treat* relations on the Web

Analysis of Dailymed indications showed that most frequent linguistic patterns indicative of the *may_treat* relation all contained the word “indicated”. This word “indicated” becomes the entry point into web data. The following steps are performed with for each drug to build a drug corpus (set of sentences) for it.

1. Use the Bing API to find the top 20 Web documents[14] with the query "druglabel" AND "indicated".
2. For each Web document:
    a. Retrieve the text from the pages (using JSoup[15]).
    b. Split the text into sentences (using Stanford Parser).
    c. Filter sentences that are too long as they will become problematic for the parser (max is set to 500 characters).
    d. Filter the sentences to keep the ones containing both the drugLabel and the word "indicated".
    e. Remove duplicate sentences.

Step (2d) is important when working with web data as there is much redundancy, and if use statistical techniques, it will affect our results.

Table 11 shows some examples of text retrieved on the web. These sentences are often not in a state that strict linguistic analysis can be performed.

Table 11. Information about experimental data (random pairs chosen)

Term	Example sentence
capsaicin	http://www.capsaicin.co.uk/pages/1369/clinical-studies the journal of proteome research stated that prior studies have indicated that capsaicin may help fight obesity by decreasing calorie intake, shrinking fat tissue, lowering fat levels in the blood.
labetalol	http://intensivecareunit.wordpress.com/2009/04/05/labetalol/ indications labetalol is indicated for the acute management of severe hypertension associated with a normal or adequate cardiac output.
methotrexate	http://www.drugs.com/pro/methotrexate-injection.html overdosage leucovorin is indicated to diminish the toxicity and counteract the effect of inadvertently administered overdosages of methotrexate.
phentermine	http://oswaldyves.com/ phentermine is indicated only for monotherapy, the drug should not be used in combination with selective serotonin-reuptake inhibitor antidepressants.

On the corpus built from the sentences, we perform the following:

1. Find the NPs. Rather than using the full parser as in the Dailymed indications, we proceed by performing part-of-speech tagging with Stanford Parser and looking for sequences of nouns.
2. For each NP, find possible associated UMLS CUIs, and keep only the NPs to which one or more (max 5) CUIs of the "disorder" semantic group can be matched.
3. Calculate the frequency of all possible UMLS CUIs to keep the most frequent ones.

[14] Web pages from the DailyMed web site are removed to not have overlapping data.

[15] JSoup is a Java HTML Parser available at http://jsoup.org/ .

In Table 12, we show some examples of NP frequencies and links to NDF-RT pairs, with the last column indicating if the pair is found or not.

Table 12. Disorder found in web sentences, its frequency rank and its presence in NDF-RT

Drug found	Disease NP	Rank	CUI	Label	Found
C0008783	ulcer	11.0	C0041582	Ulcer	no
Cimetidine	ulcer	10.0	C0007117	Basal cell carcinoma	no
	hyperacidity	3.0	C0151713	Hyperchlorhydria	no
	hypersensitivity	2.0	C0020517	Hypersensitivity	no
	heartburn	2.0	C0018834	Heartburn	YES
C0009914	side effects	7.0	C0879626	Adverse effects	no
clonidine	side effects	7.0	C0041755	Adverse reaction to drug	no
	hypertension	3.0	C0020538	Hypertensive disease	YES
C0009079	schizophrenia	35.0	C0036341	Schizophrenia	YES
Clozapine					
C0010620	reactions	10.0	C0002792	anaphylaxis	no
Cyproheptadine	hay fever	7.0	C0018621	Hay fever	no
	skin hives	4.0	C0042109	Urticaria	YES

We perform some statistics to identify the retrieval capability of this experiment. We found that 216 drugs were part of the experimental data. Of those, there were 16 drugs for which no sentences were retrieved from the Web. For the other 200 drugs, we managed to gather an average of 20 sentences from which we extracted an average of 32 NP candidates. Among the 200 drugs, for 60 of them (35%), none of the NPs generated led to information part of the NDF-RT may_treat pairs. For the other 65%, we were able to find the correct answer at an average rank of 4.

This means for about 65% of the drugs looked at, searching on the web for 20 pages and analyzing its sentences containing the drug name and the word « indicated », we are able to retrieve information found in a *may_treat* pair of a specialised and recognized resource. This is an interesting result.

We will do the same for each disease, building a disease corpus by gathering information from the web.

4.4 How are pairs actually expressed on the Web?

In the previous experiment, we showed that via the entry point “indicated”, we are able to access some *may_treat* pairs as expressed in textual data on the web. In the present experiment, we try to discover how *may_treat* pairs are actually expressed on the Web to compare the linguistic patterns found with the ones from the DailyMed analysis.

We repeat the process from the previous Web experiment for retrieving information and launch queries on Bing API of type “drugLabel” AND “disorderLabel”. We retrieve all the sentences containing the drug and the disorder from the top 20 pages. We compile all patterns no longer than 100 characters separating the pair and calculate their frequency of occurrences among all pairs.

Table 13 shows the top patterns found[16]. This is obviously noisier than the clean DailyMed patterns. The most interesting is that we do not even find among the top 20 patterns the word “indicated”. These patterns do contain words such as “calming” or “treat” and “treatment”. Many frequent patterns, “for”, “in”, “(“, or even “ ” (space) would certainly not be useful for searching on the Web.

Table 13. Patterns extracted from *may_treat* sentences on the Web

Pattern found	Frequency	Pattern found	Frequency
for	235.0	) for	27.0
	134.0	to treat	25.0
calming	65.0	external analgesic lotion, calming	23.0
in	62.0	treatment for	19.0
(	54.0	&	16.0
(tetanus (	53.0	for the treatment of	16.0
(tetanus (tetanus	53.0	with other medications	14.0
-resistant	41.0	helps slow down and reverse the process of	14.0
in the treatment of	35.0	for treating	13.0
and	31.0	pubertal	13.0

4.5 Conclusions on Web Data

We showed that with a good entry point into the web, it is possible to find interesting data. The same way as we mentioned for dailymed not being able to judge the value of the results, we would have to show the results to a domain experts. All we can evaluate is recall on known data and that we find about 65% recall at rank 4.

As with other knowledge discovery process, we would suggest to use it to show candidates to a human to be able to evaluate the interest of the disorders retrieved and decide if they should be added or not to the resource.

In the last section, we showed how the knowledge expressed on the web “naturally” is quite different than what was found in the DailyMed resource. Patterns retrieved are very noisy containing words such as and, or even just a parenthesis. We would have to perform the experiment of searching for a drug + such a pattern to evaluate results, but based on our previous background and expertise on knowledge patterns, we are sure that patterns with such general words would not lead to good results, unless the web search is first contained thematically. For example with tools like TerminoWeb [11] we gather domain-specific corpus, and then look for patterns, so a context is set.

[16] We focus on “forward” patterns, assuming the drug occurs first in the sentence and the disorder occurs second. This was true of indications in DailyMed, but not necessarily true of occurrences of drug/disorder pairs on the Web, and we will look at backward patterns in future work.

5 Conclusions

We have shown an exploration of 3 resources (1) NDF-RT with direct access to RDF info, (2) DailyMed with an indication field in which textual information can be found an analyzed, and (3) the web at large where valuable information is also found, but hidden among large amount of noisy information.

Table 14. Comparing DailyMed and Web Data

Information	DailyMed	Web Data
Nb indication sentences	3231	4172
Nb drugs covered	2090	216
Nb drugs not in *may_treat* of NDF-RT	1241	0 (we started from NDF-RT)
Nb drugs in *may_treat* of NDF-RT	849	216
Nb disorder found	457	131 (average rank 4 among candidates)
Percentage of drugs for which a known disorder was found	53.8%	60.6%

Table 14 summarizes our findings from DailyMed and the Web. The 3231 sentences in DailyMed are the subset that we could parse (using full parsing as opposed to only part-of-speech tagging for the Web). The percentage of drugs for which a known disorder was found is of 53.8% for DailyMed (457/849) and 60.6% for the Web (131/216). We obviously are not comparing equal sets, but still, the Web result is quite interesting. It shows that with a good entry point to it, the Web can lead to a recall rate comparable to a resource such as DailyMed, in which we are certain that the indication information would lead to a *may_treat* predicate. This entry point “indicated” is valuable, and with normal bootstrap methods (as is usually suggested in knowledge discovery with linguistic patterns [12]), very noisy patterns would have been found. Nevertheless, some patterns discovered on the Web seem to have potential (“treatment”, calming”) and we should explore them in future work.

Obviously, the recall evaluation toward a resource already in RDF format does not do justice to the process if it is to be used in knowledge discovery. Nevertheless, as we are not medical experts, evaluation with recall assures us that our method is at least able to retrieve known information. Information not found is not necessarily wrong but hopefully new, and should be validated to be incorporated into a specialized resource.

The coverage problem for drug labels on the Web should be investigated, as 72% of them had all their labels with 0 hits. Although UMLS contained an average of 6 labels per drug, so many were not found as they are so specific. Some simple label modifications could work, for example by excluding the information about format and quantity from the labels.

Beyond web searches for which we need adequate labels, a persistent underlying challenge in this text analysis process is the matching of labels whether we look at web sentences or sentences from DailyMed indications. In this research, we have relied on the simplest matching algorithm of Lucene, but we should investigate this further. Also, in complement to better label matching, we can explore inferencing. We sometimes saw that a more general disorder was mentioned in the text but that

NDF-RT had a *may_treat* pair with a more specific disorder. The generic-specific predicate could be used for inferences.

New experimentations should also be done on the opposite findings of drugs for known disorders.

In conclusion, much future work is envisaged to further exploit the content of each resource with more refined methods. First, we should deploy more precise analysis for parsing and distilling the knowledge from DailyMed. For the Web, we need to access quality information, and deal with noise and redundancy. Redundancy is an issue we briefly mentioned, but need to come back to. Pure copy of information is noise in a statistical process, but redundancy could be used as certainty evaluation on new information if different recognized web sources all corroborate that same information. The present research has shown that textual data found on the Web can be valuable, so it is worth exploring to provide ways of enriching specialized resources.

References

1. Rubin, D. L., Moreira, D. A., Kanjamala P. P., Musen M. A. (2008), AAAI Spring Symposium Series, Symbiotic Relationships between Semantic Web and Knowledge Engineering, Stanford University.
2. Lincoln, M.J., et al. (2004) U.S. Department of Veterans Affairs Enterprise Reference Terminology strategic overview, Studies In Health Technology And Informatics, 107, 391-395.
3. Carter J.S., Brown S.H., Erlbaum M.S., Gregg W., Elkin P.L., Speroff T., Mark S. Tuttle, M.S. (2002), Initializing the VA medication reference terminology using UMLS Metathesaurus co-occurrences. Proceedings of AMIA Annual Symposium, Boston, p.116–20.
4. Bodenreider, O. (2004), The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research 2004, vol . 32 (Database issue).
5. Jentzsch A, Zhao J, Hassanzadeh O (2009), Linking Open Drug Data, Triplification Challenge.
6. Manning, C.D, P. Raghavan, H. Schütze. Cambridge UP (2008). Classical and web information retrieval systems: algorithms, mathematical foundations and practical issues.
7. Klein D., Manning C.D. (2003), Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.
8. Miwa M., Pyysalo S., Hara T., Tsujii J. (2010), A Comparative Study of Syntactic Parsers for Event Extraction, Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, Uppsala, Sweden, p.37-45.
9. Kim J.-D., Wang Y., Takagi T., Yonezawa A. (2011), Overview of Genia Event Task in BioNLP Shared Task 2011, Proceedings of BioNLP Shared Task 2011, Workshop 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, p.7-15.
10. Auger, A. and Barrière, C. (2010) Probing Semantic Relations: Exploration and identification in specialized texts, Benjamins Current Topics, John Benjamins.
11. Agbago, A., Barrière, C. (2005) Corpus Construction for Terminology, Proceedings from the Corpus Linguistics Conference Series, Birmingham, UK.
12. Brin, S. (1998) Extracting Patterns and Relations from the World Wide Web, Proceedings of the International Workshop on the Web and Databases, pp. 172-183.