The ICD-10 Explorer: An Application of Life Science Data in SharePoint

White Paper

Abstract:

The life sciences industry adopts Semantic Web technology to realize novel applications that reduce product development costs, leverage collaboration in research or improve regulatory submissions. The "ICD-10 explorer" presented in this paper is a Microsoft SharePoint-based application to explore the standard classification of diseases. We demonstrate that Semantic Web technologies help to realize this specific application.

GRASP, the SharePoint add-on from DIQA that is used to realize the ICD-10 explorer, makes Semantic Data accessible for a number of applications in SharePoint, e.g. document retrieval, information dashboards, or semantic data integration.

Authors:
Dr. Michael Erdmann
Daniel Hansch

Pfinztalstraße 90
76227 Karlsruhe, Germany
http://www.diqa-pm.com

Introduction

Microsoft SharePoint is a Web application platform that has historically been associated with intranet content management and document management, but today it serves many other purposes as well. As an application platform it supports solutions for asset-management, collaboration, enterprise search, and business intelligence with built-in features like workflow-support, governances and security controls[1]. Special purpose solutions are found in the life sciences industry where SharePoint is e.g. used to maintain electronic health records (electronic health record system), to manage and document laboratory experiments (electronic lab notebook) and to manage clinical trials (electronic clinical trial management system)[2].

The life sciences industry is currently undergoing a transformation that is driven by (i) the need to collaborate with development partners and competitors ("coopetition"), and (ii) requirements for interoperability of data, for example for submissions to the regulatory bodies. There are various approaches to adopt Semantic Web standards in the life sciences to tackle these needs, e.g. (i) OpenPhacts tries to enhance the state of the art in drug discovery research [oPHACTS 12] and (ii) CDISC2RDF works on translating the CDISC standard into RDF[3].

Since SharePoint does not support Semantic Web standards out of the box, we have developed GRASP ("Graph for SharePoint"[4]). With GRASP data sources from the Web of Data become accessible and ontologies and taxonomies using Semantic Web standards become manageable in SharePoint, e.g. to re-use existing schemas present in the Web of Data. To demonstrate the benefits of Semantic Web technologies in the life sciences domain we have created the ICD-10 Explorer on top of GRASP as a sample application.

After giving an overview of the basic standards, ICD-10, SKOS, and the SharePoint term store, we will explain how we have translated ICD-10 from its original ClaML-XML format to SKOS and how we deployed this model. Then we will demonstrate how this knowledge can be put to good use, before illustrating further use-cases of semantic technologies and linked data for SharePoint.

[1] Microsoft Corporation: Microsoft SharePoint 2010 Evaluation Guide.
[2] For an overview of current applications with SharePoint in the life science industry: http://www.microsoft.com/health/en-us/solutions/pages/life-sciences.aspx
[3] CDISC2RDF is an initiative to publish CDISC standards in RDF: http://cdisc2rdf.com/
[4] http://diqa-pm.com/en/GRASP

Basic Standards

ICD-10

The life sciences domain has a lot of taxonomies and controlled vocabularies. A widely used standard is the International Statistical Classification of Diseases and Related Health Problems (ICD) that contains more than 10.000 terms classifying diseases and other health problems.

It is published annually by the WHO and is available in a number of variations and languages. In Germany physicians and clinics are obliged by law to encode their diagnoses according to ICD-10-GM 2013 (the German Modification of the ICD-10). "In addition to enabling the storage and retrieval of diagnostic information for clinical, epidemiological and quality purposes, these records also provide the basis for the compilation of national mortality and morbidity statistics by WHO Member States."[5]

ICD-10 contains codes and preferred names for every ICD-10 class. Additionally ICD-10 also provides hints for physicians, how they should apply the classification system, e.g. the *inclusiva* and *exclusiva*, i.e. when and when not to use a class, respectively. An example for the information contained in ICD-10-GM is given below:

- X - Krankheiten des Atmungssystems
  - J09-J18 - Grippe und Pneumonie
    - J09 - Grippe durch bestimmte nachgewiesene Influenzaviren
      - Inclusion: Influenza A/H5N1 Epidemie [Vogelgrippe]
      - Exclusion: Meningitis [G00.0] durch Haemophilus influenzae [H. influenzae]

The ICD-10 data is published in either a flat ASCII file or in the ClaML XML dialect [ClaML 06].

Simple Knowledge Organization System (SKOS)

The Simple Knowledge Organization System [SKOS 09] is the W3C standard format for representing terminologies and taxonomies. It is based on the RDF datamodel [RDF] and its semantics is well-defined. Thus, all objects in SKOS are RDF resources, all SKOS statements can be expressed as RDF triples, and all SKOS models can be represented in any of the RDF serialization syntax variants. Further, SKOS is easily extensible, by defining new properties or classes refining the properties and classes specified in the standard SKOS vocabulary.

Concepts

[5] http://www.who.int/classifications/icd/en/

In contrast to OWL-classes, which represent sets of individuals, SKOS-concepts merely represent an arbitrary concept in the world. SKOS defines some taxonomic relationships between concepts: skos:broader and skos:narrower, e.g. apple-skos:broader-fruit, europe-skos:narrower-germany. As these examples demonstrate, the interpretation of broader/narrower is not necessarily a generalization/specialization relationship as we know from OWL. They can also represent partonomic relations and others. The SKOS vocabulary also defines non-taxonomic relationships between concepts e.g. skos:related. All these relationships can be refined by extensions to SKOS.

Terms

SKOS concepts represent the things in the world that can be named. In SKOS the names of things are called terms and there are essentially two kinds of terms. The preferred name of a concept is the skos:prefLabel and all other names are alternative labels (skos:altLabel). Each concept can have at most one preferred label for each language. In case this rather simple model of concepts with labels is not sufficient and terms should be treated as objects as well, e.g. to define relationships between individual terms, SKOS-XL, the SKOS eXtension for Labels could be used. All properties defined for labeling concepts can also be refined by extensions to SKOS.

SharePoint Term Store

A functionality of SharePoint is its Managed Metadata feature[6]. It allows users to create terms with synonyms and to organize them in a number of taxonomies, so called term sets. This information is stored in the term store and can be used for annotating documents and list items. The predefined terms help SharePoint users to stick to the common language recommended or enforced by information architects. A well-defined vocabulary also has many advantages when it comes to content-retrieval. Search queries can be formulated easier, and with the help of stored synonyms the recall will be higher. Since terms are organized in hierarchies, documents tagged with sub-terms can be returned or the user can drill-down and refine the search along the lines of the term hierarchy.

From ICD-10 to SKOS and from SKOS to SharePoint

In order to provide the information from the ICD-10 to SharePoint users we employ GRASP, our semantic infrastructure for SharePoint, in particular the GRASP features to interpret SKOS files and import them into the SharePoint term store.

[6] http://technet.microsoft.com/en-US/library/ee424402%28v=office.14%29.aspx

To prepare the actual translation from ClaML-XML to SKOS via an XSLT stylesheet[7], we refined the SKOS model with ICD-10-specific extensions, e.g. subclasses for skos:Concept and subproperties for skos:prefLabel, altLabel, note and related.

There are three different types of ICD-10 classes: chapters, blocks and categories. For each of them we created a new subclass of skos:Concept, which are populated via XSLT.

To identify the concepts in the semantic web, we must create URIs for them. Since the code of ICD-10 classes is unique we create URIs by appending the code to a common prefix, which is based on the version of the ClaML file. For ICD-10GM 2013 (that was published on September 21, 2012), e.g.:

<http://diqa-pm.com/ontologies/ICD-10_2012-09-21#J09>

The common prefix also identifies the SKOS scheme, which embraces all extracted triples.

In ClaML, taxonomic relations are expressed via SubClass and SuperClass tags. The XSLT stylesheet translates them into skos:broader triples. ClaML categories can also be extended by so-called modifiers. For the SKOS model we create subcategories for each modification. The SKOS concept representing the ClaML category is linked to each subcategory via a new subproperty of skos:narrower.

In CLaML classes can have a preferred label, as well as a long and a short preferred label. SKOS, on the other hand allows exactly one preferred label (per language). Thus, we keep only the preferred label and use subproperties of skos:altLabel to represent the long and short forms.

For search applications based on ICD-10 it is important to retrieve the right ICD-10 classes also via the symptons or diagnoses identified. Thus, we chose to model the *inclusiva* (which in the ClaML file are present as free-text only) as a special form of skos:altLabel. The *exclusiva*, on the other hand are translated into a special skos:note subproperty (storing the textual information) and a special skos:related subproperty storing the relation between the current class and the excluded ones.

For many ICD-10 classes the ClaML file contains more textual information, stored in so-calles "rubrics". Most of the rubrics are translated into special subproperties of skos:note, e.g. for coding-hints, introductory texts or for definitions.

Overview of the ICD-10 extensions to SKOS, that we have developed:

[7] http://www.w3.org/TR/2007/REC-xslt20-20070123/

- Subclasses of skos:Concept: Chapter, Block, Category, SubCategory
- Subproperties of skos:altLabel: preferredLong, preferredShort, inlusion, code
- Subproperties os skos:note: text, exclusionNote, codingHint, introduction, note, modifierLink, definition
- Subproperties of skos:related: exclusion
- Subpropeties of skos:narrower: modifiedBy

Now, that all relevant information is stored in a SKOS model, we can use GRASP's built-in feature to load the SKOS file into the SharePoint term store. From now on, all ICD-10 classes with their synonyms are available within SharePoint for tagging and searching.

Use Cases for ICD-10 in SharePoint

The ICD-10 Explorer

We have developed an ICD-10 web part (Microsoft's term for SharePoint widgets) that presents textual information about a selected ICD-10 class within SharePoint pages. It is displayed in the right hand side of the user interface and displays the code and preferred label of a class, some textual information like coding hints, notes or definitions, as well as its *inclusiva* and *exclusiva* (cf. Fig. 1**Fehler! Verweisquelle konnte nicht gefunden werden.**).

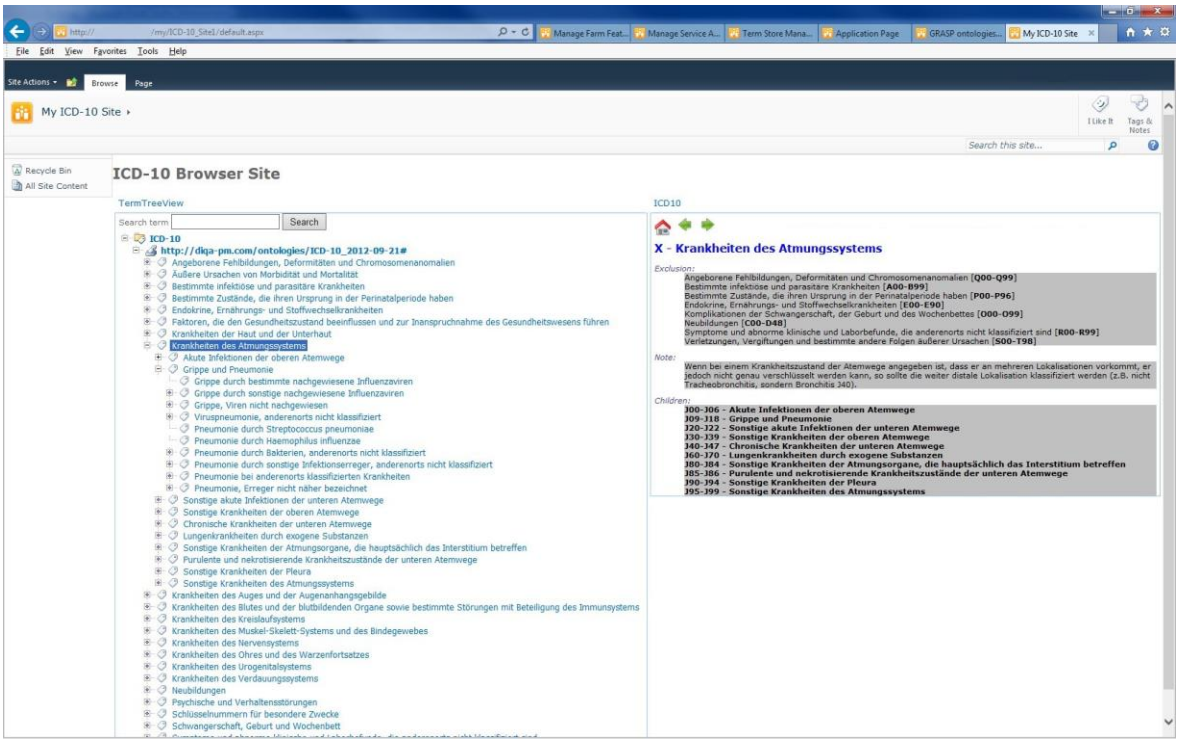

Fig. 1: The ICD-10 Explorer

On the left hand side of the screenshot we see the hierarchy of ICD-10 terms. This is a generic component provided by GRASP (the term tree view web part) that displays a particular excerpt from SharePoint's term store. Users can search terms by label or synonym. Whichever term is selected in the tree will be presented in the ICD-10 web part, thus enabling SharePoint users to browse and look-up ICD-10 codes.

Document Retrieval

Metadata is a key-feature of Microsoft SharePoint. Information architects can define table-like schemas for lists and libraries. The columns can have many different types and SharePoint will take care that any constraints coming with these types are considered, e.g. while data entry. A special column type is the so-called "managed metadata column" (cf. Figure 2). The values for managed metadata columns must come from the term store. While defining such a column architects can choose a particular term set from the term store or even select a root term. When users add documents (or other list items) to lists that use this column they will be presented with a user interface that simplifies the process of choosing appropriate (and valid) terms. In this way the information architect can make sure that users stick to the pre-defined controlled vocabulary. Having this kind of control over the entered data increases data quality and utility.

In our case only terms from the ICD-10 vocabulary are allowed. When entering values, the user could start typing "Vogelgrippe". This is part of the preferred label for

> U69.21 (Influenza A/H5N1 Epidemie [Vogelgrippe])

but also an *inclusiva* for

> J09 (Grippe durch bestimmte nachgewiesene Influenzaviren)

Since *inclusiva* were modeled as alternate labels in the SKOS model SharePoint will present J09 in the list of choices, thus guiding the user to the right code.

Once documents are properly tagged with meta-data from the term store, document retrieval via search will be much more precise. Instead of searching for all documents mentioning "Grippe" users can search for all documents tagged with "J09". Since terms are organized in a hierarchy one can even retrieve all documents in a block or chapter, e.g. all documents tagged with "J09-J18 Grippe und Pneumonie" or any of its subterms. Even exploring the search results via drill-down becomes possible, e.g. starting from the set of all documents in the block J09-J18 one can navigate to the documents tagged with J09.

All these search features can be freely combined with all other search capabilities that SharePoint has to offer, e.g. free-text search, or restrictions by document type, author, date etc.

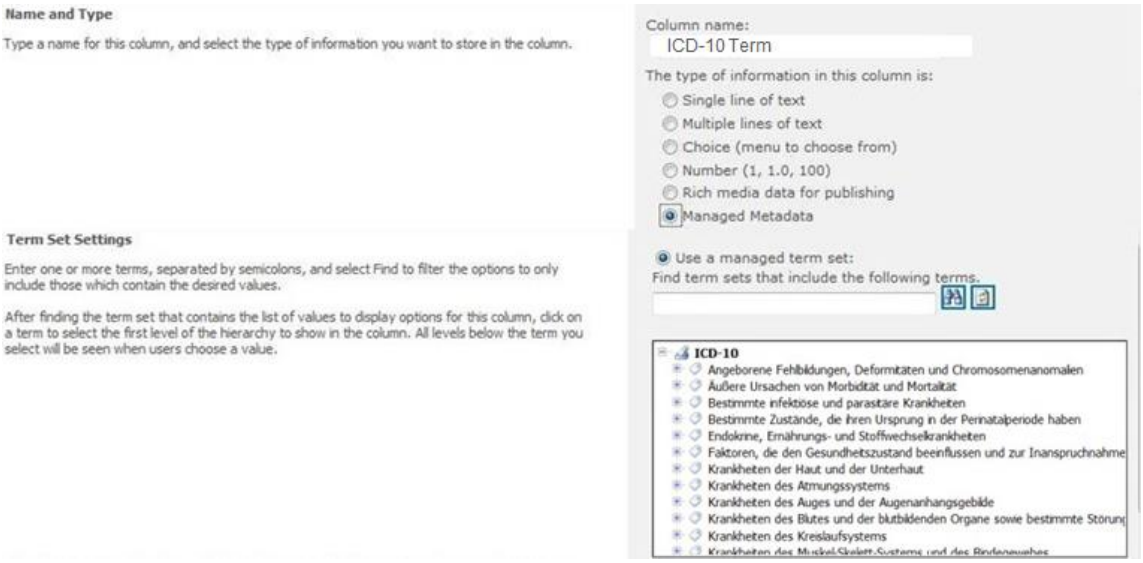

Figure 2: Creating a site-column that uses ICD-10 terms from the term store.

Using Linked Data in SharePoint

Today, many data sources are available that follow linked data principles and can be consumed and embedded in own applications. This is particularly true for the domain of life sciences, which represents a large subset of the linked data cloud[8]. With GRASP, SharePoint users now can create *information dashboards* that tap the web of data. This visualization can take different shapes, e.g. as a table or list or (when numeric data is relevant) in the form of different kinds of diagrams (e.g. line diagrams, or bar and pie chart).

One of the main strengths of the web of data is its distributed character. Different authorities provide certain data independently of each other. Due to the open standards used for publishing, anyone can use the data sources to create "mash-ups" to yield additional views and insights. GRASP makes it possible to even access multiple data sources and to combine them in one query in order to provide SharePoint users with a consolidated view. In Figure 3 we show an example SharePoint page, requesting the definition of all sub-

[8] http://lod-cloud.net/versions/2011-09-19/lod-cloud_colored.html

terms of "Catalytic Activity" from the "Gene Ontology"[9] as it is stored in the LinkedLifeData dataset[10].

Subclasses of 'Catalytic Activity' from Gene Ontology

label	definition
catalysis of free radical formation	"Catalysis of a reaction that generates a free radical, a highly reactive molecule with an unsatisfied electron valence pair." [GOC:jl]
cyclase activity	"Catalysis of a ring closure reaction." [ISBN:0198547684]
deaminase activity	"Catalysis of the removal of an amino group from a substrate, producing ammonia (NH3)." [GOC:jl]
demethylase activity	"Catalysis of the removal of a methyl group from a substrate." [GOC:mah]
first spliceosomal transesterification activity	"Catalysis of the first transesterification reaction of spliceosomal mRNA splicing. The intron branch site adenosine is the nucleophile attacking the 5' splice site, resulting in cleavage at this position. In cis splicing, this is the step that forms a lariat structure of the intron RNA, while it is still joined to the 3' exon." [GOC:krc, ISBN:0879695897]
glycogen debranching enzyme activity	"Catalysis of the cleavage of branch points in branched glycogen polymers." [ISBN:0198506732]
glyoxalase III	"Catalysis of the reaction: methylglyoxal + H2O = D-lactate." [MetaCyc:GLYOXIII-

Figure 3: Query result of a SPARQL query within a SharePoint page

As descibed in Section 4.2 users can combine imported terminologies, data from SPARQL endpoints and local SharePoint data. In the screenshot of Figure 4 we see (1) a hierarchical representation of the GeneOntology that has been imported into the SharePoint term store and (2) a list of documents that is compiled based on the selected term from the tree view, some query results from an external SPARQL endpoint and the SharePoint annotations of locally stored documents. This table shows an icon and the document title with a link to the document in the first column, the uploader and the upload date in the next columns, followed by the terms the documents have been tagged with. All this data is stored in SharePoint. The last column titled "Definition" shows the definition of the associated with the tags that is retrieved from the SPARQL endpoint.

[9] http://geneontology.org/
[10] http://linkedlifedata.com/

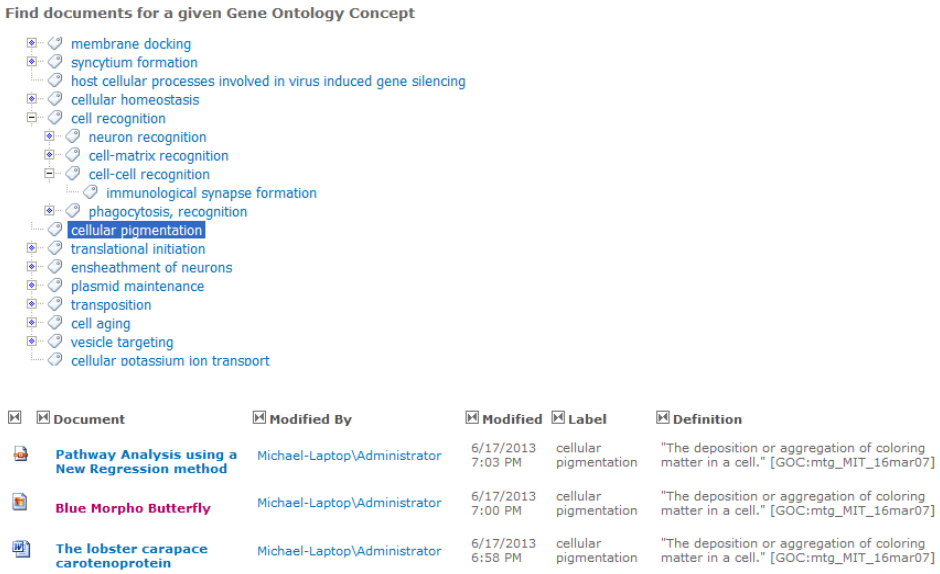

Figure 4: List of documents that is created by combining local SharePoint data with queries to the web of data

Summary

In this white paper we have demonstrated how semantic technologies can be employed in SharePoint. Giving an example workflow for transforming ICD-10 first in a semantic knowledge representation format and then into SharePoint's native term store we were able to demonstrate the benefits and also point to advanced, further use cases of semantic technologies that become possible with GRASP.

With GRASP, DIQA has created the tools for SharePoint information architects to easily create applications that benefit from the Web of Data to provide SharePoint users with relevant data for their decision making processes.

The use cases supported by GRASP include

- **Ontology management** (incl. rights management and version control)
- **Taxonomy and terminology management** (reusing SKOS taxonomies and exporting term sets to an open format, supports multiple parents, provides "working copies" of terminologies that can be edited in SharePoint)
- **Queries and visualization** of local ontologies or of remote SPARQL endpoints
- **Data integration** between data from different SPARQL sources

If you want to find out more about GRASP please visit the GRASP homepage at http://diqa-pm.com/en/GRASP. There you can also find a trial version for evaluation purposes and contact information, if you have additional questions.

References

[ClaML 06] Egbert J. van der Haring, S. Broënhorst, Huib ten Napel, S. Weber, M. Schopen, Pieter E. Zanstra: ClaML: A standard for the electronic publication of classification coding schemes. In proceeding of: Ubiquity: Technologies for Better Health in Aging Societies - Proceedings of MIE2006, The XXst International Congress of the European Federation for Medical Informatics, Maastricht, The Netherlands, August 27-30, 2006. pp.801-806

[GRASP 12] Michael Erdmann, Daniel Hansch: Applications of the Web of Data in SharePoint. White Paper. DIQA Projektmanagement GmbH 2012. http://diqa-pm.com/en/DIQA_unlocks_the_Web_of_Data_for_SharePoint_users

[oPHACTS 12] Williams A., Harland L., Groth P,. et al.: Open PHACTS: Semantic interoperability for drug discovery. Drug Discovery Today, June 06, 2012

[SKOS 09] Alistair Miles, Sean Bechhofer: SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009. http://www.w3.org/TR/skos-reference/

[SP2010] Microsoft Corporation: Microsoft SharePoint 2010 Evaluation Guide. http://download.microsoft.com/download/0/B/0/0B06C453-8F7D-4D8E-A5E5-D50DC6F8D8F4/SharePoint_2010_Evaluation_Guide.pdf

[SPARQL 12] Steve Harris, Andy Seaborne: SPARQL 1.1 Query Language. W3C Proposed Recommendation 08 November 2012. http://www.w3.org/TR/sparql11-query/