# Challenges in Collecting Patient Characteristics for Analysis using Linked Data

*B. Kämpgen[1], K. März[2], M. Hafezi[4], A. Adler[1], P. Philipp[1], S. Zelzer[3], A. Saffari[4], M. Nolden[3], A. Mehrabi[4], L. Maier-Hein[2], A. Rettinger[1]*

*[1] AIFB, Karlsruhe Institute of Technology, Karlsruhe, Germany*
*[2] Junior group Computer-assisted Intervention, DKFZ, Heidelberg, Germany*
*[3] Division of Medical and Biological Informatics, DKFZ, Heidelberg, Germany*
*[4] Visceral and Transplantation Surgery, University of Heidelberg*

Kontakt: kaempgen@kit.edu, k.maerz@dkfz.de

***Abstract:***
*Decision support systems in medical research require a both specialized and flexible knowledge representation to accommodate for a wide range of use cases. In this short research paper we describe the challenges encountered when designing a suitable representation for patient characteristics. We present a light-weight RDF ontology, allow medical personnel to describe and annotate liver factors effectively, and explain why collaborative modelling proved difficult.*

*Keywords: Decision Support, Knowledge Modeling*

## 1 Problem Statement

A main barrier to create efficient decision support systems is the integration of patient data from heterogeneous data sources [1]. The Transregional Collaborative Research Centre (TCRC) "Cognition-Guided Surgery"[1] envisions a cognitive system that perceives data, interprets it using a knowledge base and performs suitable actions, thus providing synergy effects to research groups working on a variety of fields ranging from radiotherapy to heart and liver surgery.
However, there are various problems to building a project-wide knowledge base: Identification and collection of relevant patient characteristics, convenient automatic or User-Interface-based [1] annotation of patients according to these factors and evolution of data and knowledge are examples for such problems. The fact that medical and technical personnel have vastly different requirements in terms of functionality and usability complicates the matter even more.
For example, in liver surgery planning there are hundreds of factors to consider, including tumor types, gen analyses and the biography or family history of a patient. Knowledge about these liver factors changes constantly, new ones are discovered and the importance of others is revised as medical research progresses. This makes it hard to create a knowledge base that is both structured enough to be useful to all participants and flexible enough for constant change.
In this short research paper, we provide the following contributions:
(1) We introduce a light-weight ontology for representing patient factors in a highly structured and flexible manner, using the example of liver surgery. We used the vocabulary of the Resource Description Framework (RDF) which facilitates data exchange on the web by exploiting semantic descriptions.
(2) We describe the architecture that allows medical personnel to enter patient data quickly and effectively in a familiar environment and serve it to other participants as well as transform and modify it.
(3) We describe the lessons learned from creating these solutions in a project with highly heterogeneous personnel.
After explaining our approach in Section 2, we discuss our results in Section 3, describe related work in Section 4 and conclude in Section 5.

## 2 Material and Methods

Fig. 1 illustrates our approach of collecting patient data in the Knowledge Base. Applications access data from the Knowledge Base via the Linked Data principles[2], i.e., URIs for naming things, the HTTP protocol for accessing data, and RDF for information representation.
The Knowledge Base consists of two main components only accessible by collaborators: 1) Surgipedia, a collaborative and lightweight ontology development tool based on Semantic MediaWiki (SMW) that publishes structured information as Linked Data. 2) The extensible Neuroimaging Archive Toolkit (XNAT), a software platform designed to facilitate

[1] http://www.cognitionguidedsurgery.de/
[2] http://www.w3.org/DesignIssues/LinkedData.html

multi-center based research, is used as common data infrastructure for sharing data between collaborating projects. Data in the Knowledge Base needs to be stored in either Surgipedia or XNAT or be published as Linked Data and reachable via links from other data in the knowledge base.

## 2.1. Describing Liver Factors

See (1) Fig. 1 for our approach to collect liver factors. One senior, two junior and three student physicians identified by experience and from literature relevant liver factors in an Original Liver Factor Table using Microsoft Excel. Excel is installed on many clinical workstations, is familiar to most physicians and provides flexibility in changing structure and appearance of the liver factors. However, deriving a structure for liver factors and storing them in a database for structured queries is difficult. The original liver factor table contained 395 liver factors in 20 worksheets. Table structure was made for medical personnel and is difficult to understand for technicians. A mapping from a row to a single liver factor was not always possible due to general names such as "Other", additional facets distinguishing between liver factors and a varying number of classification levels of liver factors. Furthermore, references to other sources (literature, SNOMED, Wikipedia, etc.) were not provided, semi-structured information was found in footnotes, various background information, e.g., data type, unit, normal range were unspecified and changing.

**Modelling using Ontology of Liver Factor Tables and Excel:** We have collaboratively created a light-weight OWL ontology to describe liver factor tables. The ontology reuses the RDF Data Cube Vocabulary (QB)[3], a W3C-standardised vocabulary for multidimensional datasets.It allows to describe datasets (`sp:Liver_Factor_Table` subclass of `qb:DataSet`) of usages (`sp:Liver_Factor_Observation` subclasses of `qb:Observation`) of liver factors (`sp:Liver_Factor`). Note that the ontology allows to both describe *empty* and *filled-in* liver factor tables. See Fig. 2 for an example of a liver factor observation.

We distinguish three namespaces: `sp` for the properties and categories of the liver factor table ontology defined in Surgipedia; `lft-version` for pointing to an empty liver factor table with a specific version uploaded to XNAT and `lft-patient` for pointing to a specific patient in XNAT.

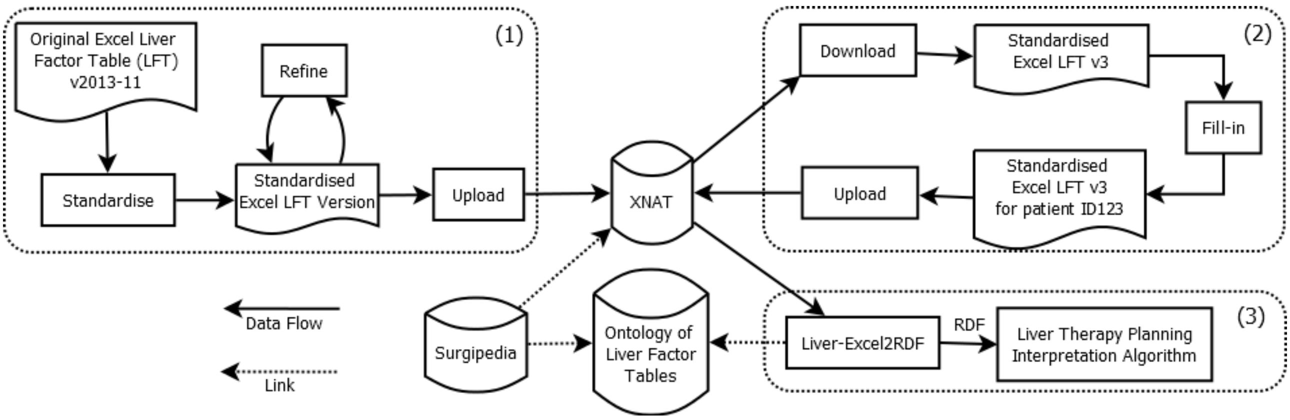


**Fig. 1**. Flow diagram illustrating approach of describing liver factors (1), annotating patient records (2), and consuming data from knowledge base (3)

Every liver factor observation has a value for a *dependent* measure `sp:obsValue` and values for *independent* dimensions relevant for interpreting the measure value. For empty liver factor tables the liver factor table itself, the liver factor, the unit, and possibly additional *feature characteristics* independent from the liver factor, e.g., `lft-version:excessive_nitrogen_load` are relevant. Liver factors are categorized hierarchically, e.g. "Disease of the central nervous system" (CNS) is more general than "Encephalopathy".

Since the original Liver Factor Table was insufficiently structured to allow automatic transformation from Excel to RDF, we have manually transformed the original liver factor table to a *Standardized Liver Factor Table* with a well-defined structure, and enable physicians to refine liver factors within this standardized version. During the automatic transformation *Liver-Excel2RDF* uses an Excel macro and *Spread2RDF*[4]. After a physician has uploaded a filled-in liver factor table to XNAT, applications such as liver therapy planning algorithms (see (3) in Fig. 1) can load it into an RDF store and use SPARQL 1.1 queries to work with the data, e.g. to automatically enrich liver factors with background information.

[3]http://www.w3.org/TR/vocab-data-cube/
[4]https://github.com/marcelotto/spread2rdf

## 2.2. Annotating patients

See (2) Fig. 1 for our approach to enable physicians to annotate liver factors for patients in the knowledge base. In our project, physicians volunteered to go through paper-based patient folders from more than 300 patients. It stands to reason that tools for efficient annotation must be supplied.

**Using standardized Excel Tables**: Given the standardized Liver Factor Tables and the Liver Factor Table Ontology we have two possibilities to provide annotation interfaces for physicians: Short term, physicians download standardized Liver Factor Table versions from XNAT, fill in copies of the table, and upload filled-in a Liver Factor Table for each patient to XNAT. Consuming applications use Liver-Excel2RDF to translate patient annotations to RDF reusing the Liver Factor Table Ontology and the version of the reused standardized Liver Factor Table and load the data into an RDF store for queries, e.g., to automatically derive new information. To continue annotating a patient, physicians download the filled-in liver factor table from XNAT, refine the annotations and update the file in XNAT. Unfortunately, this solution does not solve the versioning problem since new versions of the standardized Liver Factor Table cannot be used to refine existing annotations that were created using older versions.

**Automatic Generation of Input Forms**: Therefore, long term, we plan to use the RDF representations of liver factor tables to automatically create user-friendly forms. To this end we will leverage prior work from the Common Toolkit (CTK)[5], a community effort to provide unified GUI elements, a standardized plugin framework and DICOM support for medical imaging applications. One CTK module that we extended builds upon the Slicer Execution Model and its command line module parameter description[6]. It was unified with an XML schema describing a parameter set and a C++ library that can generate a GUI for editing this parameter set. After linking and consolidating [3] different liver factor table versions and automatically creating a partly-filled-in form, physicians can refine annotated patients after updates to the liver factor table.

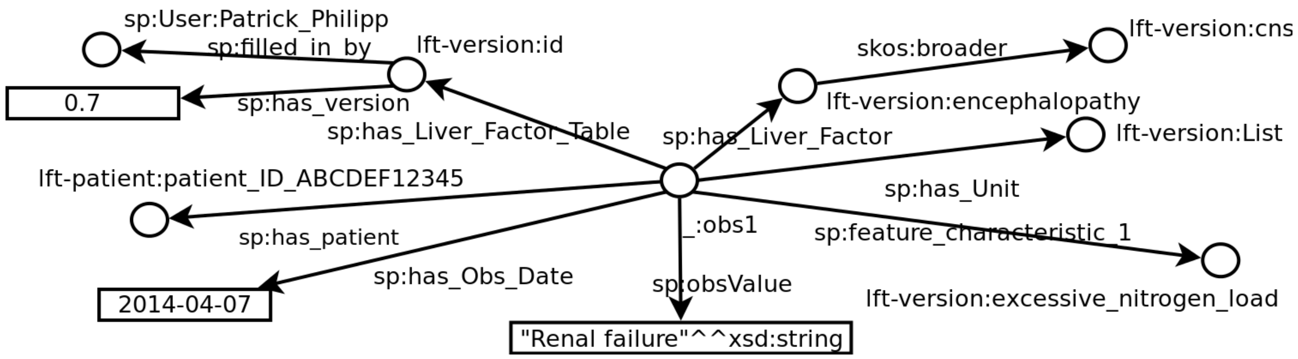


**Fig. 2.** Flow diagram illustrating approach of describing liver factors (1), annotating patient records (2), and consuming data from knowledge base (3)

# 3 Lessons Learned

Collaborative modelling and annotation of liver factors using a semantic wiki did not meet the needs of both technicians and physicians. The first version of the liver factor table was written in Microsoft Excel. Since Excel is lacking collaboration support, we have tried to introduce Surgipedia for collaboratively describing liver factors. We imported the standardized Liver Factor Table into Surgipedia and tutored physicians in the use of SMW. Some problems could be solved this way, e.g., time series could be modeled as so-called subobjects in SMW. Also, categorization using “subpropertyOf” was possible so that all liver factors could be visualized as a tree or graph. Although SMW fulfilled most of the modelling requirements from a technical point of view, the solution was not applicable: The editing user interface was too complicated and cumbersome for physicians. The support for collaboration and discussions was not used. As a result, some physicians continued working on the Excel versions of Liver Factor Tables. Merging changes from Excel and Surgipedia was costly.

Similarly, we tried to apply Surgipedia as a liver patient annotation tool. We imported the liver factor tables by enriching Excel-sheets with Wiki syntax for templates and forms, exporting such Excel-sheets as HTML and importing the resulting HTML to a Surgipedia test server. There, we had five student physicians that in total annotated 20 patients using forms. This solution had the advantage that the Excel-to-HTML-export resulted in visualizations closely-resembling the original Liver Factor Table and still providing additional benefits such as auto-completion on input fields. One could re-

[5] http://www.commontk.org/docs/html/index.html

[6] http://www.commontk.org/docs/html/CommandLineModules_Page.html

use the built-in RDF export functionality of SMW. Also, Inline Queries directly on the structured data allowed overviews in tables, the calculation of the average age of annotated patients and integrity constraint checks [5]. Still, this solution was not applicable: Significant manual effort is necessary when applying changes of new versions of liver factor tables to their wiki representation. Also, it is difficult to apply Access Control Lists to wiki systems, as is necessary for sensitive patient data.

# 4 Discussion

Designing a multi-project knowledge base is difficult in regard to making the system accessible to both physicians and researchers. Our solution allows medical personnel to enter data quickly in a familiar environment, while making the data available in the RDF for further computation. Medical personnel on the project are now switching to the new Excel sheets for primary case documentation of the case, which shows that the method is viable.

We distinguish general approaches for capturing, representing and using Electronic Patient Data and specific approaches using Semantic Web concepts for data modelling and integration. González-Ferrer et al. [1] describe barriers to creating an efficient Clinical Decision Support System (CDSS): integrating patient data from heterogeneous systems and devices. They come to the conclusion that the HL7 Virtual Medical Record (vMR) fulfills front-end requirements such as user-friendliness and openEHR archetypes based on vMR fulfil back-end requirements such as expressivity. We argue that using Linked Data for representing and sharing of patient data fulfils those requirements and provides better interoperability between heterogeneous systems [3].

Leroux and Lefort [4] also use the RDF Data Cube Vocabulary to represent patient data. They describe how to automatically map clinical trial data to standards (e.g. UMLS, SMOMED-CT, LOINC) and show how to add information about sensors collecting trial data using the Semantic Sensor Network Ontology. In contrast to our approach, they do not deal with the difficulties of allowing physicians to define and annotate patient factors. Related to our planned work on automatic creation of annotation interfaces is RaUL [2], a vocabulary for web forms. The CTK XML specification we intend to use promises better support for healthcare specificities such as security, images and time series data.

# 5 Summary

In this work, we present a solution to collect liver factors and annotations thereof for patients in a knowledge base for a large interdisciplinary project. To this end, we have introduced a lightweight ontology for describing versioned Liver Factor Tables. We expect that our approach can be applied to define and collect patient factors from other domains such as heart surgery. An alternative collaborative approach for liver factor collection and annotation proved unsuccessful due to the characteristics of the healthcare domain. In future work, we plan to thoroughly evaluate our approach according to a first liver surgery planning interpretation algorithm using annotated liver factors from the knowledge base.

# 6 Acknowledgments

This work was carried out with the support of the German Research Foundation (DFG) as part of projects I01 and A02, SFB/TRR 125 Cognition-Guided Surgery.

# 7 References

[1] González-ferrer, A., Peleg, M., Verhees, B., Verlinden, J.m.: *Data Integration for Clinical Decision Support Based on openEHR Archetypes and HL7 Virtual Medical Record*. In: Process Support and Knowledge Representation in Health Care, Lecture Notes in Computer Science, **vol. 7738**, pp. 71-84. Springer Berlin Heidelberg (2013)

[2] Haller, A., Hausenblas, M.: *RaUL : RDFa User Interface Language A Data Processing Model for Web Applications*. In: Chen, L., Triantafillou, P., Suel, T. (eds.) Web Information Systems Engineering WISE 2010, Lecture Notes in Computer Science, **vol. 6488**, pp. 400-410. Springer Berlin Heidelberg (2010)

[3] Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A., Decker, S.: *Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine*. Web Semantics: Science, Services and Agents on the World Wide Web 9, 365-401 (2011)

[4] Leroux, H., Lefort, L.: *Using CDISC ODM and the RDF Data Cube for the Semantic Enrichment of Longitudinal Clinical Trial Data*. ict.csiro.au (2012), http://www.ict.csiro.au/staff/laurent.lefort/swat4ls2012.pdf

[5] Vrandečić, D.: Ontology evaluation. Springer (2009)