

# ADT2FHIR – A Tool for Converting ADT/GEKID Oncology Data to HL7 FHIR Resources

Noemi DEPPENWIESE<sup>a,b,1,\*</sup>, Pierre DELPY<sup>c,d,\*</sup>, Mohamed LAMBARKI<sup>c,d</sup> and  
Martin LABLANS<sup>c,d</sup>

<sup>a</sup>Medical Center for Information and Communication Technology,  
Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg  
(FAU), Erlangen, Germany

<sup>b</sup>Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg  
(FAU), Erlangen, Germany

<sup>c</sup>Federated Information Systems, German Cancer Research Center (DKFZ),  
Heidelberg, Germany

<sup>d</sup>Complex Data Processing in Medical Informatics, University Medical Centre  
Mannheim, Mannheim, Germany

**Abstract.** Harmonized and interoperable data management is a core requirement for federated infrastructures in clinical research. Institutions participating in such infrastructures often have to invest large degrees of time and resources in implementing necessary data integration processes to convert their local data to the required target structure. If the data is already available in an alternative shared data structure, the transformation from source to the desired target structure can be implemented once and then be distributed to all participants to reduce effort and harmonize results. The HL7® FHIR® standard is used as a basis for the shared data model of several medical consortia like DKTK and GBA. It is based on so-called resources which can be represented in XML. Oncological data in German university hospitals is commonly available in the ADT/GEKID format. From this common basis we conceptualized and implemented a transformation which accepts ADT/GEKID XML files and returns FHIR resources. We identified several problems with using the general ADT/GEKID structure in federated research infrastructures, as well as some possible pitfalls relating to the FHIR need for resource ids and focus on semantic coding which differs from the approach in the ADT/GEKID standard. To facilitate participation in federated infrastructures, we propose the ADT2FHIR transformation tool for partners with oncological data in the ADT/GEKID format.

**Keywords.** Data Management, Medical Oncology, Health Information Interoperability, FHIR, Open Source

<sup>1</sup> Corresponding author: Noemi Deppenwiese, Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), noemi.deppenwiese@fau.de

\* Equal contribution as first author: Noemi Deppenwiese and Pierre Delpy

## 1. Introduction

Federated medical consortia like the German Cancer Consortium (DKTK) or the German Biobank Alliance (GBA) aim at interconnecting university hospitals and biobanks respectively across Germany. One goal of such consortia is to establish federated data managements for their various information on patients, studies and biosamples across all participating sites (biobanks [1], university hospitals [2,3]). To support harmonization, both GBA [4] and DKTK [5,6] have recently modeled their sample-related and oncology-focused datasets in FHIR. This FHIR-based data can be used for example in federated search infrastructures, as is currently the case in the *bbmri.de* Sample Locator which was developed by GBA. Other projects like the Medical Informatics Initiative and NUM Codex also use FHIR-based data models for search and data analysis [7]. The ADT/GEKID dataset [8] is a standard in Germany mandated by law for reporting oncology-related data to cancer registries. In order to make it easier for participating sites in consortia like GBA or DKTK to integrate their oncology data into the FHIR stores, we intended to provide them with a proven tool to transform ADT/GEKID XML files into FHIR resources.

## 2. Methods

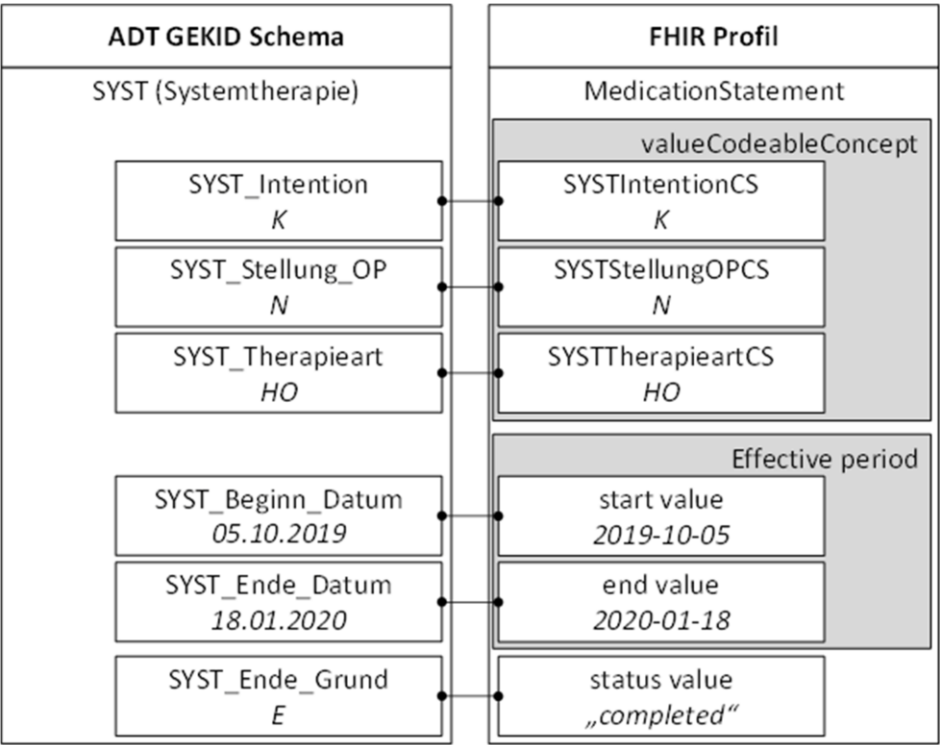
We decided to base our implementation on XSLT, since it is lightweight, open source, an open standard and both ADT/GEKID and FHIR resources can be represented in XML.. Therefore, it can be easily integrated into existing ETL infrastructures as well as open-source tools (e.g. DKTK-Bridgehead). Because of its original purpose (reporting information about cancer cases for statistical purposes), the ADT/GEKID format is focused on providing information about the respective current state of the patient for several key dates (e.g. diagnosis, begin of treatment or end of treatment). In contrast to this report-centered snapshot view, the FHIR data model can also incorporate information on past or future (planned) events, e.g. a former diagnosis that is no longer active or a planned procedure. For this work, we used the FHIR representation from DKTK and the Oncology Centers of Excellence [5,6] which to our knowledge is currently the only complete and up-to-date FHIR representation of the ADT/GEKID dataset.

The DKTK focuses on a modified, research-oriented view of the ADT/GEKID dataset, the so-called “Meldedatensatz“ (MDS), which is less centered around the report entity and instead focuses on the events (e.g. diagnoses or therapies). Due to its focus on medical research, it contains additional data elements (e.g. sample, consent) and is better suited as a basis for a representation in FHIR. We therefore decided to approach the transformation in two steps: First transform the ADT/GEKID XML into the MDS XML format (ADT2MDS) and then transform the MDS XML into FHIR XML (MDS2FHIR).

As a prerequisite for the transformation into MDS, an extensive analysis and mapping of the ADT/GEKID scheme was performed. ADT2MDS was originally developed in an initiative funded by the German Cancer Aid to extend the DKTK IT network to German comprehensive cancer centers (Connecting Comprehensive Cancer Center, C<sup>4</sup>). To cover additional data elements, a specific extension of the official ADT/GEKID\_v2.1.1 XSD scheme was used. In a first step, all theoretical implementations of the data structure were analyzed and generated test data was taken into account. In the second step we used test data from C<sup>4</sup> sites. The last step iteratively

integrated feedback after a first rollout. To validate technical operation, the transformation was applied on real patient data at C<sup>4</sup> sites.

In FHIR, resources can either be communicated as separate entities or grouped together in FHIR bundles, which is a kind of container resource. Since storing all generated resources for one ADT/GEKID file in one bundle might result in impracticably large bundles, one bundle will be created per patient. In the MDS view, all information on a single patient is already stored under one XML node, so one patient node corresponds to one FHIR Bundle in the transformation output.



**Figure 1.** Schematic example of an ADT/GEKID-SYST-Entity and the corresponding FHIR-MedicationStatement illustrating data input and output of the ADT2FHIR tool.

While the DKTK FHIR Implementation Guide [5] uses many code systems identical to those used in the ADT/GEKID XML, in some cases a mapping was necessary. First, in ADT/GEKID the reasons for stopping a radiation or systemic therapy are communicated with a special code. In FHIR medication (systemic therapy) or procedure (radiation therapy) resources, this information can be represented using the *status* attribute. Since the values of *status* are limited to a set given in the FHIR specification, it is necessary to map the ADT/GEKID values. The values *E* and *R* correspond to the *completed* status, while *U* results in an *unknown* status. An example for such a mapping for the *E* status can be seen in figure 1. All other codes point to a *stopped* status. If no stopping reason code is given, the status is either *active* or *intended* for systemic or *in-progress* or *preparation* for radiation therapy, depending on the start date. The second

case was information about sample material. While there is information about samples in the ADT/GEKID data set, the corresponding DKTK Implementation Guide does not specify a profile for samples, encouraging implementers to use the profile from the GBA core data set Implementation Guide [4] instead. Therefore, it became necessary to map the codes for *fixation type*, *sample type* and *sample material* to the GBA *sample material type*. This mapping is displayed in tables 1 and 2.

**Table 1.** Mapping ADT sample material (“Probentyp”) codes to GBA sample material type codes for “Probentyp = ‘Flüssigprobe’ ” (fluid samples).

ADT sample material	GBA sample material type
Vollblut	whole-blood
Serum	blood-serum
Plasma	blood-plasma
Urin	urine
Liquor	csf-liquor
Knochenmark	bone-marrow
DNA	dna
RNA	rna
Protein	derivative-other
<i>Other values</i>	liquid-other

**Table 2.** Mapping ADT sample material (“Probentyp”) and fixation type (“Fixierungsart”) codes to GBA sample material type codes for “Probentyp = ‘Gewebeprobe’ ” (tissue samples).

ADT sample material	ADT fixation type	GBA sample material type
Tumorgewebe	Kryo/Frisch (FF)	tumor-tissue-frozen
Normalgewebe	Kryo/Frisch (FF)	normal-tissue-frozen
<i>Other values</i>	Kryo/Frisch (FF)	other-tissue-frozen
Tumorgewebe	Paraffin (FFPE)	tumor-tissue-ffpe
Normalgewebe	Paraffin (FFPE)	normal-tissue-ffpe
<i>Other values</i>	Paraffin (FFPE)	other-tissue-ffpe

### 3. Results

The transformation tool is split into two XSLT files, available under <https://github.com/samplify/ad2fhir>, which can be executed sequentially to generate multiple files, each containing a FHIR bundle with all data related to one patient, from a single ADT XML file.

For the ADT2MDS transformation, extensive analysis and validation was necessary. As a result, we determined that the ADT/GEKID scheme is highly tolerant in its restrictions. In theory, this leads to infinite representations of possible ADT/GEKID xml files. For instance, identifiers for patients, diagnoses or any other tumor related entity are defined as optional. In practice, however, analysis of test data we received from C<sup>4</sup> sites clearly demonstrated a less ambiguous application of the ADT/GEKID scheme in actual tumor documentation systems. At least patient and diagnosis identifiers were always applied in all tested cases. Based on test data generated from the extended ADT/GEKID scheme as well as the test data provided from sites, we wrote, validated and improved our ADT2MDS transformation. The feedback from the C<sup>4</sup> real patient data testing revealed additional interpretations of the extended ADT/GEKID scheme which we then integrated in our transformation. However, this work is still in progress and we expect further feedback to be integrated.

We realized that using the original MDS model could in some cases lead to data loss and adapted the transformation script for our purposes. For instance, the ADT/GEKID entities of course (*Verlauf*), operation, radiation therapy and system therapy are represented in a summary “progress” entity in MDS. We recovered the information loss by re-adding those ADT/GEKID elements and then mapped the MDS objects to FHIR resources, using the existing DKTK FHIR Implementation Guide as a reference.

The output files of the second transformation step were then validated against the profiles of the DKTK FHIR Implementation Guide, showing no errors. Additionally, we manually validated our transformation results by comparing data elements of anonymized ADT/GEKID real patient data to the FHIR output data.

One particular problem of the transformation was the handling of FHIR resource ids. While the ADT/GEKID specification does not enforce ids on any level (ids for entities like diagnosis, operation etc. are optional), FHIR requires an id for each resource. XSLT provides a function for obtaining a unique id per node called *generate-id*, which we used to generate ids for resources where the corresponding nodes had none. In cases where more than one resource is generated from one node, these ids are extended using date values or custom strings in order to ensure each resource has a unique id.

In order to use the ADT/GEKID id of a patient also as a business identifier for the patient in the corresponding FHIR patient resource (not just as a technical resource id), this identifier also needs a system. Since this is a local Identifier for the patient, it also needs to be a local system URL. Therefore, the second script should be adapted by each site and the *http://fhir.example.org/LokaleTumorPatientenIds* be replaced with a meaningful one in order to ensure unique identifier values for the system.

In some cases, ADT/GEKID allows free-text values for attributes that become CodeableConcepts in FHIR. In these cases, the values were copied into the *CodeableConcept.text* value, with the *CodeableConcept.coding* fields staying empty.

#### 4. Discussion

We implemented a transformation from ADT/GEKID XML files to FHIR resources. We were able to successfully test the transformation with several example data sets, including real patient data from several sites, producing valid FHIR resources in accordance with the DKTK FHIR Implementation Guide.

Since the ADT/GEKID scheme allows a wide variety of interpretations, our ADT2MDS transformation only covers a subset of those, in particular the ones which occurred in the validation steps. However, this applies only to ADT2MDS, since the structure of the MDS output is well defined. An additional limitation of this ADT2MDS transformation is that it only applies to the ADT/GEKID base dataset and does not account for organ specific modules.

The developed transformation uses the ids from the ADT/GEKID file as FHIR resource ids. If these ids are not yet sufficiently pseudonymized, this might lead to privacy issues in some environments. In this case, pseudonymization or anonymization either of the input ADT/GEKID file or the generated FHIR resources may be necessary.

The key issue we identified is the possibility of missing ids. If an updated ADT/GEKID file is processed, new resources will be created for all contained data elements. This is complicated by the lack of required ids in ADT/GEKID. The *generate-id* function, used to obtain ids for resources without corresponding ADT/GEKID ids, will create different ids for the same entities if a different XSLT implementation is used or

the document structure changes. Therefore, the consistency of ids for the FHIR resources cannot be relied on. This means that, if not all necessary ids are set, updated data requires either removing all previously imported data or some sort of post processing to identify resources that are (updated) duplicates of existing resources. This pitfall not only applies to ADT/GEKID, but to all cases where XML-based data is to be converted to the realm of FHIR without required id elements for all entities that correspond to a separate FHIR resource. It could be mitigated by modifying the ADT/GEKID scheme to require ids for all entities corresponding to a FHIR resource. However, in practice most optional ADT/GEKID entity ids are filled in.

If a biobank already has FHIR resources for all its samples, the resources generated for the samples from the ADT/GEKID file might be duplicates and need to be removed to avoid counting these samples more than once.

Another problem is the difference in consistency of terminology usage. ADT/GEKID elements sometimes allow free text where coded data would be possible. For example, ADT/GEKID allows to specify substances used in systemic therapy as free text. In contrast, FHIR heavily relies on semantically coded data. As described before, while this data can be stored in the *CodeableConcept.text*, if ADT/GEKID would support semantic codes in these elements, the values could be stored in *CodeableConcept.coding* instead, greatly improving interoperability, especially if a standardized terminology is used. For future development, we aim to include additional organ specific modules from the ADT/GEKID dataset as well as profiling those additional data elements in FHIR. Also, it would be feasible to include an option to hash all ids in order to avoid privacy problems. The Medical Informatics Initiative is currently also developing a FHIR-based core oncology dataset which is also influenced by ADT/GEKID and builds on the same “oncology” DTK profiles [6]. Once this work is finished, we plan to update the ADT2FHIR tool to support the resulting profiles.

In General, the generation of FHIR resources based on ADT/GEKID enables interoperability between clinical data holders. On the one hand, they can easily exchange data on a technical level, while using FHIR functionalities. On the other hand, all tools that are built on that FHIR data can be reused and further be developed, since they work on the same data model.

## Declarations

*Conflict of Interest:* The authors declare that they have no conflict of interest.

*Contributions of the authors:* ND, PD: developed the transformation and wrote the manuscript; MLam: specification and mapping of the oncology FHIR dataset; MLab: conceived and supervised the project and provided input on the manuscript. All authors have approved the manuscript as submitted and accept responsibility for the scientific integrity of the work.

The present work was performed in (partial) fulfillment of the requirements for obtaining the degree “Dr. rer. biol. hum.” from the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (ND).

The present work was funded by the German Cancer Consortium and the German Biobank Alliance.

## References

- [1] C. Schüttler, N. Buschhüter, C. Döllinger, L. Ebert, M. Hummel, J. Linde, H.-U. Prokosch, R. Proynova, and M. Lablans, Anforderungen an eine standortübergreifende Biobanken-IT-Infrastruktur: Erhebung des Stakeholderinputs zum Aufbau eines Biobanknetzwerks der German Biobank Alliance (GBA), *Pathologe*. **39** (2018) 289–296. doi:10.1007/s00292-018-0435-9.
- [2] S. Joos, D.M. Nettelbeck, A. Reil-Held, K. Engelmann, A. Moosmann, A. Eggert, W. Hiddemann, M. Krause, C. Peters, M. Schuler, K. Schulze-Osthoff, H. Serve, W. Wick, J. Puchta, and M. Baumann, German Cancer Consortium (DKTK) – A national consortium for translational cancer research, *Mol Oncol*. **13** (2019) 535–542. doi:10.1002/1878-0261.12430.
- [3] M. Lablans, E.E. Schmidt, and F. Ückert, An Architecture for Translational Cancer Research As Exemplified by the German Cancer Consortium, *JCO Clinical Cancer Informatics*. (2018) 1–8. doi:10.1200/CCI.17.00062.
- [4] bbmri.de, BBMRI.de/GBA Implementation Guide, (2021). <https://samplify.github.io/bbmri-fhir-ig/> (accessed March 29, 2021).
- [5] M. Lambarki, Datenmodell des DKTK und der Onkologischen Spitzenzentren, (n.d.). <https://simplifier.net/oncology/> (accessed March 29, 2021).
- [6] M. Lambarki, J. Kern, D. Croft, C. Engels, N. Deppenwiese, A. Kerscher, A. Kiel, S. Palm, and M. Lablans, Oncology on FHIR: A data model for distributed cancer research, *Stud Health Technol Inform.* (accepted).
- [7] L.A. Kapsner, M.O. Kampf, S.A. Seuchter, J. Gruendner, C. Gulden, S. Mate, J.M. Mang, C. Schüttler, N. Deppenwiese, L. Krause, D. Zöller, J. Balig, T. Fuchs, P. Fischer, C. Haverkamp, M. Holderried, G. Mayer, H. Stenzhorn, A. Stolnicu, M. Storck, H. Storf, J. Zohner, O. Kohlbacher, A. Strzelczyk, J. Schüttler, T. Acker, M. Boeker, U.X. Kaisers, H.A. Kestler, and H.-U. Prokosch, Reduced Rate of Inpatient Hospital Admissions in 18 German University Hospitals During the COVID-19 Lockdown, *Front. Public Health*. **8** (2021) 594117. doi:10.3389/fpubh.2020.594117.
- [8] J. Dudeck, W. Wächter, U. Altmann, and F. Katz, The definition of a new uniform basic data set for hospital cancer registries in Germany., *MIE-Proceedings*. (1993) 489–92.