German Medical Data Sciences: Bringing Data to Life
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210070

Oncology on FHIR:A Data Model for Distributed Cancer Research

Mohamed LAMBARKI[a,1], Jori KERN[a], David CROFT[a], Cäcilia ENGELS[d], Noemi DEPPENWIESE[f], Alexander KERSCHER[c], Alexander KIEL[e], Stefan PALM[b], and Martin LABLANS[a,g]

[a]*Federated Information Systems, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[b]*West German Cancer Center, University Hospital Essen, Essen, Germany*
[c]*Comprehensive Cancer Center, University of Würzburg, Würzburg, Germany*
[d]*Charité University Medicine Berlin, German Biobank Node, Berlin, Germany*
[e]*Leipzig Research Centre for Civilization Diseases, University of Leipzig, Leipzig, Germany*
[f] *Chair of Medical Informatics, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany*
[g] *Federated Information Systems, University Medical Center Mannheim, Mannheim, Germany*

Abstract. In the field of oncology, a close integration of cancer research and patient care is indispensable. Although an exchange of data between health care providers and other institutions such as cancer registries has already been established in Germany, it does not take advantage of internationally coordinated health data standards. Translational cancer research would also benefit from such standards in the context of secondary data use. This paper employs use cases from the German Cancer Consortium (DKTK) to show how this gap can be closed using a harmonised FHIR-based data model, and how to apply it to an existing federated data platform.

Keywords. FHIR, Oncology, Interoperability, DKTK, biobanking, comprehensive cancer center

1. Introduction

Large scale medical research often spans dozens of institutions, making it imperative that the data being collected and shared is comparable across sites. To this end, data handling systems need to be able to exchange harmonised data across locations and organisations. These systems in turn need well-defined health data standards to enable interoperability. At present, HL7's Fast Healthcare Interoperability Resources (FHIR) promises to address the interoperability problem, improving usability of health data [1]. Established in 2012 as the federated data management body of the German Cancer Consortium (DKTK) [2], the Clinical Communication Platform (CCP) plays a central role in the collection and exchange of clinical data and biosamples among German cancer centres

[1] Mohamed Lambarki, German Cancer Research Center (DKFZ), Division of Federated Information Systems, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany; E-mail: m.lambarki@dkfz-heidelberg.de

and thus provides an important bridge between disciplines and institutions [3]. To this end, the CCP uses the so-called bridgehead to connect the multiple participating sites of the consortium. With its distributed infrastructure, consisting of multiple data warehouses and search interfaces, the bridgehead enables scientists to query and request data and biosamples across partner institutions [4]. As a result of the expansion of the bridgehead architecture to all Comprehensive Cancer Centers (CCC) and biobanks of the German Biobank Node [5], maintaining interoperability among the bridgeheads in over 22 university hospitals is becoming a major issue. This stems from the fact that data transformation from one dataset to another can induce a loss of information [6]. Also, serving multiple research networks (each with their own focus on disease areas or types of data or biomaterial) by hosting multiple instances of the bridgehead software would lead to added overhead for local IT operations. Instead, we propose to create a modular FHIR dataset that covers oncology but can be expanded to different use-cases. The modular nature of such a dataset allows a common data format to be used in data transfer between multiple sites, and makes it possible to store the data in a structured manner.

2. Methods

Since all German cancer care providers are obligated to report patient data to cancer registries following the ADT/GEKID schema [7], we used its XML representation as a starting point and compared it with the extended version created in DKTK at the data element level and the structural level. In order to simplify the identification of the entity types for the data model, a codification of the cancer disease (cancer lifecycle) was created based on the analysis of the tumor documentation (Figure 1). The next step was to determine how to represent the DKTK data model using HL7 FHIR. Based on the previous analysis, the FHIR resources were identified and the structure of the model was visualised with the clinFHIR software [8]. The results were discussed with technical and domain experts in various meetings and telephone conferences. Before expanding and adapting the FHIR resources (profiling) to fit various DKTK and ADT data elements, other oncology FHIR profiling endeavours were analysed regarding their reuse in DKTK: (1) mCode, an initiative of the American Society of Clinical Oncology (ASCO) in collaboration with the MITRE Corporation, aims to establish a core set of structured data elements for oncological electronic health records [9]. (2) Breast Cancer Data (BCD), a joint project of the Clinical Information Council (CIC) and the Clinical Information Modeling Initiative (CIMI), maintains a set of data elements used for breast cancer staging [10]. (3) The Cancer Care Record (CCR), initiated at the University Hospital of Cologne, form the basis for the dataset of the national Network Genomic Medicine (nNGM) Lung Cancer [11]. Another project, which was originally designed for the data transmission of ADT in CDR format, also provides a solid basis as a guide for the modelling of cancer data[12]. The following tools were used to implement the FHIR profiles: (1) Forge: FHIR profile editor, a desktop application for tailoring the FHIR resources to specific FHIR profiles [13]. (2) Simplifier: HL7 FHIR registry for hosting FHIR profiles and examples [14]. (3) FHIR-Validator: Java application to validate FHIR profiles and examples [15]. The definition of the profiles was created based on the FHIR specification in its current version FHIR R4 [16]. We also used Logical Observation Identifiers Names and Codes (LOINC) [17]. To speed up the profiling process, an additional Java program was developed, which parses the ADT schema and the DKTK Metadata Repository [18] (Samply MDR, implementing the ISO

11179 standard). It generates collections of permissible terms that contain coded values of data elements (FHIR ValueSets and CodeSystems). In addition, it supplements the FHIR profiles with extra information about the data elements, such as the cardinality, "must-support" markers and a human-readable description of the content. To take into account the habits of the user and domain experts and to help them understand the data model and the mapping between the specific data content of the FHIR resource, the DKTK data set (in its former implementation) and the ADT specification, the individual data elements and the mapping to FHIR were provided as an "MDRSheet" Excel spreadsheet [19]. In some cases, there were no corresponding data elements in the FHIR specification. These cases were evaluated to determine whether additional FHIR extensions were necessary. For each profile, a corresponding example was created. These examples were packaged together into a single FHIR bundle [16]. For the FHIR modelling of certain data structures such as the TNM Classification of Malignant Tumours (TNM) [20], several viable alternative models could be constructed. To build the one best suited to DKTK purposes, criteria such as simplicity and compatibility with the data set of the German cancer registries were applied.

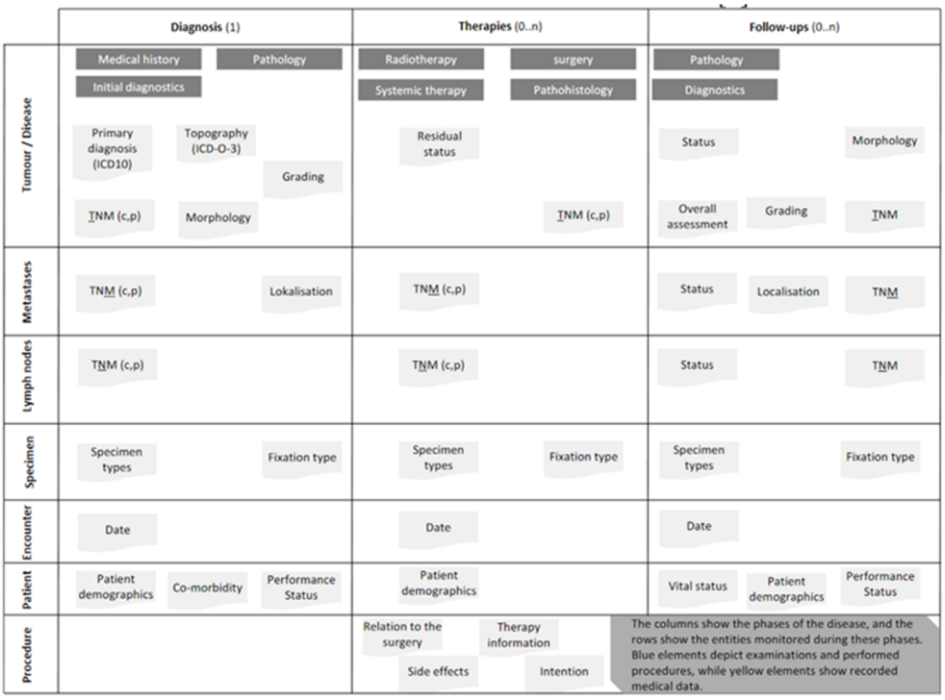

Figure 1. Lifecycle of tumour documentation (simplified).

3. Results

FHIR is based on the concept of "resources" describing many aspects of healthcare [21]. The following FHIR resources were identified as being relevant to the current DKTK data model: Condition, Observation, Procedure, MedicationStatement, Patient, Organization, Specimen, ClinicalImpression and Encounter. As part of our analysis, we

performed a search for existing oncology FHIR profiles, looking in particular for ones that might already meet our requirements. However, we were not able to find a perfect match. The primary issues can be summarized as follows: (1) FHIR version: all profiles of the projects analysed (mCode, CCR, BCD) were specified with old FHIR versions (STU2/3). (2) Value sets: value sets were defined in SNOMED, which was, at the time of our analysis, not yet licensed in Germany (mCode). (3) Modelling: certain clinical information was too detailed and thus distributed over multiple FHIR profiles. This makes the data both difficult to read and to retrieve (mCode, BCD). (4) Specificity: some profiles were organ-specific, describing e.g. only breast cancer (BCD).

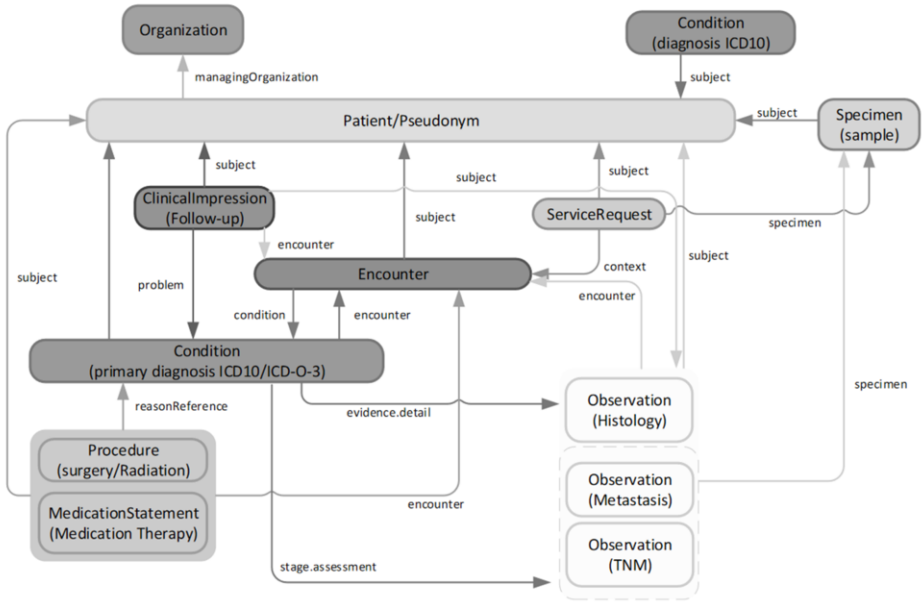

Figure 2: FHIR data model: Resources (simplified). Colours indicate different resource types, arrows represent the references between the resources.

3.1. Data Model

The content of the data model in its current version covers the use cases of DKTK and, to describe biosample data, the German Biobank Alliance (GBA) [22]. In order to ensure the compatibility and possible data ingress from the established tumour documentation, the architecture is based on the ADT data structure. The data model is also designed to be: (1) simple and realistic, reflecting clinical and biomedical reality, (2) flexible and expandable for specialised applications, (3) future-proof andinteroperable. In addition to the technical requirements, functional considerations werealso taken into account. The main requirements for the model can be summarized asfollows: (1) Allow direct linking of all events with the patient (examinations, therapiesetc.). (2) Representation of the primary diagnosis through the combination of ICD-10[23] and ICD-O-3 [24] (localization/topography). (3) Additional linkage of all events(sample, therapies etc.) directly with the primary diagnosis (as far as possible). (4) Chronological representation of the different events (e.g. examinations and therapies). (5) Consideration of other diseases the patient may be suffering from (e.g. comorbidities). (6) Coverage of different

therapy types in relation to the primary diagnosis. (7) Representation of follow-ups and further development of the primary tumour, the metastases and the lymph nodes.

In its current version, the data model of the DKTK and the Comprehensive Cancer Centres [25] (Figure 2) comprises three categories of information: Patient-identifying data, medical/clinical data and data about biosamples. The clinical data include the primary diagnosis, treatments and the subsequent follow-ups. The data model allows biosample data to be recorded dependent or independent of the primary diagnosis.

3.2. FHIR Profiles

In the following, the relevant profiles are described according to the categories already mentioned (see section 3.1). Detailed descriptions can be viewed on the Simplifier platform. Of particular note in this context are the following points: (1) no elements have been removed from the base FHIR Resources, in order to maintain interoperability with other FHIR implementations. (2) All file names and URLs of the FHIR profiles were named according to a consistent naming scheme.

3.3. Clinical data: primary diagnosis

For each primary tumour, several data elements are required for a complete description. These include the classification according to ICD-10 and ICD-O-3, as well as body-site-laterality. The primary diagnosis can reference further Observations (TNM, histology) as evidence and assessment for the primary diagnosis (The words between the brackets show the underlying base FHIR resource):

- **Primary diagnosis** (Condition): includes parameters used to classify the primarydisease such as ICD-10, ICD-O-3 (topography) and the body site, as well as links to histology (topography) and TNM classifications (pathological/clinical)
- **Histology** (Observation): indicates the histology of the tumour, based on the current ICD-O-3 classification (morphology)
- **Metastasis** (Observation): indicates whether metastases have been found
- **Grading** (Observation): indicates the degree of tumour differentiation
- **TNMc / TNMp** (Observation): clinical or pathological staging of the primary tumour according to the current TNM classification of UICC

3.3.1. Clinical data: therapy

For each primary diagnosis, the therapies performed (surgery, radiation, systemic therapies) are defined using the following profiles. The therapy outcome (R-classification) is incorporated into the model via further Observations:

- **Surgery** (Procedure): includes surgical data such as OPS [26] and body site
- **Radiotherapy** (Procedure): covers the radiotherapy data
- **Systemic therapy** (MedicationStatement): includes chemotherapy and all types of therapy not covered by surgical data or radiotherapy
- **Residual status** (Observations): overall and local assessment of the residual tumour (R-classification)
- **Overall assessment of tumour status** (Observation): overall assessment of the disease, such as partial remission, complete remission and progressive disease

3.3.2. Clinical data: follow-ups

As with the primary diagnosis, the follow-up is linked to several observations, such as the status of the tumour, the lymph nodes and the metastases:

- **Follow-up** (ClinicalImpression): contains parameters for the development of the disease and is linked to corresponding Observations, especially TNM and histology. ClinicalImpression was chosen as a counterpart to the "progression" in ADT following a discussion in the HL7 interoperability forum. This resource would serve as a kind of data container with a series of observations for tracking the progression of the primary disease.
- **Treatment case** (Encounter): models a timeline of the clinical progression of the disease. This includes different clinical events such as therapies and observations.
- **Vital status** (Observation): records the vital status of the patient such as date of last contact or death

3.3.3. Biosamples

Depending on the use case, the biosample data can be linked to the primary diagnosis and/or the patient. The biosample data includes in particular the type and date of sampling. This profile was taken with minor modifications from the German Biobank Alliance: Specimen (Specimen): records of biosample data. It includes, for example, the type and date of collection of the sample. This profile acts as a bridge between the profiles of oncology and the profiles of biobanks (GBA), allowing researchers to search for biosamples in biobanks associated with specific patient groups/diagnoses [27] via the bridgehead infrastructure. In order to link the biomaterial to the primary diagnosis, the resource ServiceRequest was used. Since this link has already been realised by an extension, this profile will most likely be removed in the next releases.

3.4. Terminology and Mapping

Coded values used to describe clinical concepts in the use cases under consideration can be found on the Simplifier platform. Whenever possible, the codes from the ADT dataset were used to create the FHIR ValueSets. In some cases, coded data types were missing in the ADT dataset. For these we substituted the more precise definitions from the DKTK's existing dataset. The Observation profiles, such as histology, metastasis and TNM (see section 3.3) were given unique codes (usually LOINC or OIDs). This is needed to identify the correct definition, for example, by a FHIR server.

FHIR provides a notation to allow mapping between the data elements of FHIR and other established metadata specifications [1]. This feature was used to map FHIR data elements to corresponding ADT and ISO standard 11179-3 fields.

4. Discussion

Standardising data specifications is an important step towards making patient-related information sharable across organisations [6] or – in our case – across research networks. To this end, an oncology model FHIR standards-based architecture for DKTK use cases was developed, incorporating feedback from both experts in the domain oncology and in

medical informatics. An important result of this work is that by using the HL7 FHIR standard, it is also possible to model specific clinical domains such as oncology without significant gaps. The use of embedded mapping annotations helps to bridge the gap between FHIR and the already existing national standards.

However, we also faced some difficulties during the development of the profiles:

- The ADT data set as a national reference has, for example, value sets that are missing or deviate from the international standard (e.g. the version of the ICD-10 and the body-site-laterality)
- Some observations lacked suitable LOINC codes (e.g. UICC staging, tumour topography and morphology); this could now be remedied by using SNOMED.
- For many value sets, such as the UICC TNM staging system, no national or international code systems are available. Creating code systems to fill these gaps would greatly assist in making interpretation across different health care organisations more comparable
- In Germany, SNOMED's license is restricted to the Medical Informatics Initiative (MII) [28] and, thus, not usable at every site participating in our networks. Therefore, the ubiquitous adoption of FHIR profiles with SNOMED annotations was not feasible but should be addressed in the future.

To address these problems, to bring this work to the wider community and to share the experience and results with other organisations in the healthcare in general and in the oncology domain in particular, discussions have already been held with HL7-Germany and the national MII [29]. The following steps were agreed upon:

- Combined with parts of other oncology-related FHIR profiling endeavours (section 2), this work will provide base FHIR oncology profiles for the MII
- Foundation of a community for the maintenance and support of German oncological base profiles as an open source project

While FHIR was originally designed for information exchange rather than storage, it offers a solid information model reflecting decades of international expertise and can be queried with both FHIR Search and Clinical Quality Language. Thus, we are beginning to store FHIR resources natively in the bridgehead. This also avoids complex transformations from other data models to FHIR and back.

With the words of Palfrey and Gasser, interoperability is not an end in itself, but a way to achieve higher societal goals; in this case improving the quality of healthcare [30].This work and its adoption as a standard by the bridgehead infrastructure is a small step in that direction.

5. Acknowledgements

This work was funded by the German Cancer Consortium (DKTK).

References

[1] M.L. Braunstein, Health informatics on FHIR: how HL7's new API is transforming healthcare, Springer, Cham, 2018.

[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.
[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.
[4] M. Lablans, D. Kadioglu, M. Muscholl, and F. Ückert, Exploiting Distributed, Heterogeneous and Sensitive Data Stocks while Maintaining the Owner's Data Sovereignty. Methods Inf Med **54** (2015).
[5] M. Lablans, D. Kadioglu, S. Mate, I. Leb, H.-U. Prokosch, and F. Ückert, Strategien zur Vernetzung von Biobanken. Klassifizierung verschiedener Ansätze zur Probensuche und Ausblick auf die Zukunft in der BBMRI-ERIC. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz **59**(2016), 373–378.
[6] V. Powell, F.M. Din, A. Acharya, and M.H. Torres-Urquidy, Integration of Medical and Dental Care and Patient Data, 1st Edition, Springer London, London, 2012.
[7] U. Altmann, F.R. Katz, and J. Dudeck, A reference model for clinical tumour documentation. Stud Health Technol Inform **124** (2006), 139–144.
[8] A. Walinjkar, FHIR Tools for Healthcare Interoperability. BJSTR **9** (2018).
[9] K.M. Miller R, HL7 FHIR Implementation Guide: minimal Common Oncology Data Elements (mCODE) Release 1-US Realm, http://hl7.org/fhir/us/mcode/ [cited 2020 March 23].
[10] Breast Cancer, http://hl7.org/fhir/us/breastcancer/2018Sep/profiles.html.
[11] Stefan Lang, Cancer-Care-Record, https://github.com/clinical-integration-hub/Cancer-Care-Record.
[12] IG:Übermittlung onkologischer Daten, http://wiki.hl7.de/index.php?title=IG:%C3%9Cbermittlung_onkologischer_Daten.
[13] Forge - FHIR, https://fire.ly/products/forge/ [cited 2020 March 9].
[14] Simplifier.net, https://fire.ly/products/simplifier-net/ [cited 2020 March 9].
[15] FHIR Validator, https://confluence.hl7.org/display/FHIR/Using+the+FHIR+Validator.
[16] HL7.org, Resourcelist - FHIR v4.0.1, https://www.hl7.org/fhir/resourcelist.html.
[17] S.M. Huff, R.A. Rocha, C.J. McDonald, G.J. de Moor, T. Fiers, W.D. Bidgood, A.W. Forrey, W.G.Francis, W.R. Tracy, D. Leavelle, F. Stalling, B. Griffin, P. Maloney, D. Leland, L. Charles, K. Hutchins, and J. Baenziger, Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. J Am Med Inform Assoc 5 (1998), 276–292.
[18] D. Kadioglu, B. Breil, C. Knell, M. Lablans, S. Mate, D. Schlue, H. Serve, H. Storf, F. Ückert, T.Wagner, P. Weingardt, and H.-U. Prokosch, Samply.MDR - A Metadata Repository and Its Application in Various Research Networks. Stud Health Technol Inform **253** (2018), 50–54.
[19] A.-K. Kock-Schoppenhauer, B. Kroll, M. Lambarki, H. Ulrich, S. Stahl-Toyota, J.K. Habermann, P. Duhm-Harbeck, J. Ingenerf, and M. Lablans, One Step Away from Technology but One Step Towards Domain Experts-MDRBridge: A Template-Based ISO 11179-Compliant Metadata Processing Pipeline. Methods Inf Med **58** (2019), e72-e79.
[20] P. Hermanek and L.H. Sobin, TNM Classification of Malignant Tumours, Springer Berlin Heidelberg, Berlin, Heidelberg, 1987.
[21] T. Benson and G. Grieve, Principles of Health Interoperability: SNOMED CT, HL7 and FHIR, 3rdEdition, Springer International Publishing, Cham, 2016.
[22] C. Klingler, M. von Jagwitz-Biegnitz, M.L. Hartung, M. Hummel, and C. Specht, Evaluating the German Biobank Node as Coordinating Institution of the German Biobank Alliance: Engaging with Stakeholders via Survey Research. Biopreserv Biobank (2019).
[23] D.I.f.M. Dokumentation und Information, ICD-10-GM 2020 Systematisches Verzeichnis, picturaWerbung, Lich, Hess, 2019.
[24] A. Fritz, ed., International classification of diseases for oncology: ICD-O, World Health Organization,Geneva, 2000.
[25] C.H. Brandts, Innovating the outreach of comprehensive cancer centers. Mol Oncol **13** (2019).
[26] B. Graubner, ICD und OPS. Historische Entwicklung und aktueller Stand. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz **50** (2007), 932–943.
[27] S. Mate, P. Vormstein, D. Kadioglu, R.W. Majeed, M. Lablans, H.-U. Prokosch, and H. Storf, On-The-Fly Query Translation Between i2b2 and Samply in the German Biobank Node (GBN) Prototypes. Stud Health Technol Inform **243** (2017), 42–46.
[28] SNOMED CT: Häufig gestellte Fragen | Medizininformatik-Initiative,https://www.medizininformatik-initiative.de/de/snomed-ct-haeufig-gestellte-fragen.
[29] S.C. Semler, F. Wissing, and R. Heyder, German Medical Informatics Initiative. Methods Inf Med **57**(2018), e50-e56.
[30] J. Palfrey and U. Gasser, Interop: The Promise and Perils of Highly Interconnected Systems, BasicBooks, New York, 2012.