MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

© 2022 International Medical Informatics Association (IMIA) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI220032

Improving Findability of Digital Assets in Research Data Repositories Using the W3C DCAT Vocabulary

Matthias Löbe[a], Hannes Ulrich[b], Christoph Beger[a], Theresa Bender[c], Christian Bauer[c], Ulrich Sax[c,d], Josef Ingenerf[b], Alfred Winter[a]

[a] *Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Leipzig, Germany*
[b] *Institut für Medizinische Informatik, University of Lübeck, Lübeck, Germany*
[c] *University Medical Center Göttingen, Göttingen, Germany*
[d] *Campus Institute Data Science (CIDAS) Göttingen, Germany*

Abstract

Research data management requires stable, trustworthy repositories to safeguard scientific research results. In this context, rich markup with metadata is crucial for the discoverability and interpretability of the relevant resources. SEEK is a web-based software to manage all important artifacts of a research project, including project structures, involved actors, documents and datasets. SEEK is organized along the ISA model (Investigation – Study – Assay). It offers several machine-readable serializations, including JSON and RDF. In this paper, we extend the power of RDF serialization by leveraging the W3C Data Catalog Vocabulary (DCAT). DCAT was specifically designed to improve interoperability between digital assets on the Web and enables cross-domain markup. By using community-consented gold standard vocabularies and a formal knowledge description language, findability and interoperability according to the FAIR principles are significantly improved.

Keywords:
Health Information Systems, Semantic Web, Data Management

Introduction

The recent discussions about the lack of reproducibility of scientific workflows and the reusability of research data motivate more elaborate approaches in research data management. The FAIR Guiding Principles [1] are now widely accepted and implemented in many projects. With the goal of developing a FAIR Data Infrastructure, various approaches are being developed for designing FAIR Digital Objects [2], exposing data as FAIR Data Points, or assessing a level of maturity of FAIRness [3]. This work is also particularly important because although Data Management Plans (DMPs) are part of many funding programms and the majority of scientific journals require statements about making the original data available prior to acceptance of a publication. Still specific requirements regarding preferrable repositories or quality controls regarding the posted dataset are often lacking. This leads to a multitude of digital repositories (e.g. re3data, a registry for research data repositories, already lists over 2,000 entries [4]). The sustainability of these solutions will be a challenge. And although metrics have been developed to assess the usefulness of biomedical digital repositories [5], these are hardly tracable for the individual researcher.The problem to be addressed in this paper is that many research data archives maintain only a minimal set of metadata, which on top of that is often simply embedded as text in web pages and not retrievable in a machine-readable way. These archives are mainly intended for manual use, where the user either browses through predefined navigation hierarchies or finds results by simple textual search. Even if a web API exists, application-specific vocabulary based on the archive's internal data model regularily hamper the usage. This results in a steep learning curve for potential users of the API. While this still works for small repositories that are homogeneous in content, repositories with a wide range of topics require different methods, e.g. for reasons of inconsistent terminology across different scientific fields. The research objective of the work presented here was to improve findability and interoperability by adding machine-readable annotations to research data assets in a vocabulary that fulfills five conditions: (1) widely used, (2) consented, (3) domain-relevant, (4) structured and (5) machine-readable. The research hypothesis is that the W3C Data Catalog Vocabulary (DCAT) in its current version 2 (2020) [6] is a viable candidate for achieving these goals. According to its creators, it is primarily "designed to facilitate interoperability between data catalogs." Furthermore, the GOFAIR initiative lists DCAT as one of the reference frameworks for criterion F2: Data are described with rich metadata .

The software SEEK [7] serves as the basis for our work. SEEK is originally a platform for managing systems biology data and models. It is a powerful open source software that is used in a larger number of projects, from large international initiatives such as FAIRDOMHub [8] to national networks such as NFDI4Health [9]] to institute-specific deployments [10]. SEEK builds on the ISA data model [11], which describes a hierarchy of 'Investigation' (the project context), 'Study' (a unit of research) and 'Assay' (analytical measurement). ISA is a popular pre-FAIR-era framework from the field of bioinformatics that includes a conceptual foundation, a technical specification, broad support from specialized tools, and an extensive community.

SEEK offers a human-readable HTML view with JSON-LD embedded, a machine-readable JSON API and a RDF serialization based on the JERM ontology and Dublin Core, but currently not DCAT.

As a reference for practical implementation, we use a well established research data repository, the Leipzig Health Atlas (LHA) [12]. The LHA is an infrastructure for archiving scientific results with a focus on clinical and epidemiological research, systems medicine, bioinformatics, medical informatics and biometrical modeling. Obviously, this includes very heterogeneous data types such as clinical or genomic data, derived

phenotyping, algorithmic models or ontological knowledge bases. The LHA is used to archive important project deliverables, such as dataset snapshots used in publications, but also as a data sharing platform. Access can be adjusted depending on the need to protect the personal health data it contains. During the development of the LHA, the FAIR Guiding Principles were always kept in mind, however many of the recommendations are intentionally broad and need to be made more specific by the community in each research area through developing best practices. So, when FAIR recommends that descriptions of data and metadata should be "rich", "domain-relevant', and "community-consented" with the help of "FAIR vocabularies"; serialized in a "broad applicable" language and associated with "detailed" provenance, one such best practice could be the proliferation of common vocabularies like DCAT.

Methods

In order to implement support for the DCAT standard in ISA/SEEK, we developed a mapping for all essential entities of the ISA/SEEK model to artifacts from DCAT. Since the application areas of ISA and DCAT overlap only partially, classes from other W3C vocabularies were also used, on which DCAT itself is based as well. To avoid misinterpretations regarding the mapping, we included four experts from three different locations and a consensus was subsequently reached. Evaluation of the mapping between SEEK/ISA and DCAT was done by an independent ontology expert, who did not take part in the mapping process.

In turn, a complex example of a clinical trial with different data bodies and distributions was fully modeled in DCAT. In particular, mapping gaps or extension options of the ISA/SEEK model were examined, for example, to better map semantic relationships by using typed relations available in DCAT.

The finally evaluated mapping was integrated into SEEK by extending the already existing RDF serialization functionalities. Any modifications of the SEEK source code takes place in a public fork of the official seek4science/SEEK GitHub repository and will be made available to the original code base via pull request.

Results

The following results were obtained during the work, which are explained in detail in the following sections:

1. We modeled a comprehensive DCAT example based on a real-world published research project from the Leipzig Health Atlas repository. As far as possible, an attempt was made to provide all attributes with meaningful characteristics.
2. We developed a mapping from the SEEK/ISA data model to the DCAT data model. It was investigated which existing characteristics could not be mapped.
3. We are implementing an extension in SEEK which implements the mapping. The extension is currently being evaluated in the Leipzig Health Atlas development system.

Comprehensive DCAT example

To exemplify the expressive power and coverage of DCAT and its relationship to SEEK, an existing, we modeled a typical entry from the LHA's research data management system, including the referenced resources (see Figure 1). The HNSCC study [13] is an observational cohort study that characterizes head and neck squamous cell carcinomas (HNSCC) with different HPV16 DNA and RNA (E6*I) status. The database used in the publication consists of clinical data from the hospital information system and genomic data from tumor sequencing. In order to support the reproducibility of research and to promote the reuse of data, the clinical data are offered in various data files compliant to different standards, including for example simple comma separated data or complex structures according to the CDISC ODM standard. To ensure traceability of the study, the underlying biometric model was archived as an R package, but it is only accessible upon request. The HNSCC study is part of the overall work on this topic (head and neck squamous cell carcinomas investigation) and associated with two research projects and two agents (see Figure 1 last line at the bottom).

For certain datasets, extended services are available in the LHA in addition to an access request or direct download. This concerns e.g. visualizations of data as Shiny Server Apps, calculators for practical testing of biometric models or the possibility to generate aggregated data as reports. In our example, for clinical datasets, a service for generating new research hypotheses can be accessed via the data warehouse i2b2 [14]. This service is not an original part of the SEEK platform, but its integration can be well described in DCAT, too.

Mapping from the SEEK/ISA to DCAT

The mapping of each resource of the example to classes of the DCAT standard is described in Table 1. All resources could be mapped. Since the ISA model was developed specifically for bioinformatics research, but DCAT aims at a domain-neutral description of digital assets, it is not an exact mapping, but a broad match. The example results in a general mapping rule for the classes from SEEK/ISA. For reasons of space, the mapping of the individual attributes of each class can only be given here in excerpts. Table 2 shows a compact example of the mapping rules for an assay (SEEK) into a dataset (DCAT). The asterisk denotes generic mappings that are not specific to assays/datasets (from dcat:Resource).

Implementation of DCAT mapping in Leipzig Health Atlas

SEEK uses the Ruby plugin (gem) RDF::Vocab to provide terms and relations from multiple schemes. At the time we developed the DCAT modification, SEEK used version 3.0.4 of RDF::Vocab, which is missing many terms from DCAT version 2, thus we had to update the dependency at least to version 3.0.11 as well as related dependencies.

The Serialization of objects and their properties is realized in SEEK with a CSV mapping file (comma/character separated values), where for each property of an object of SEEKs data model, respective RDF terms or relations are provided, including Ruby code to transform the property values to literals or URIs. We modified this CSV file and added new entries to reflect our mapping from JERM to DCAT. Because of the fact that we used new schemes, which originally were not supported by SEEK (i.e., DCAT and PROV), we had to include new namespace entries into SEEK's RDF serialization process.

All changes are available at https://github.com/Leipzig-Health-Atlas/seek/tree/feature/dcat-mapping and we plan to propose them to the developers of SEEK, so that they may be included into the core functionality of SEEK in a future update.

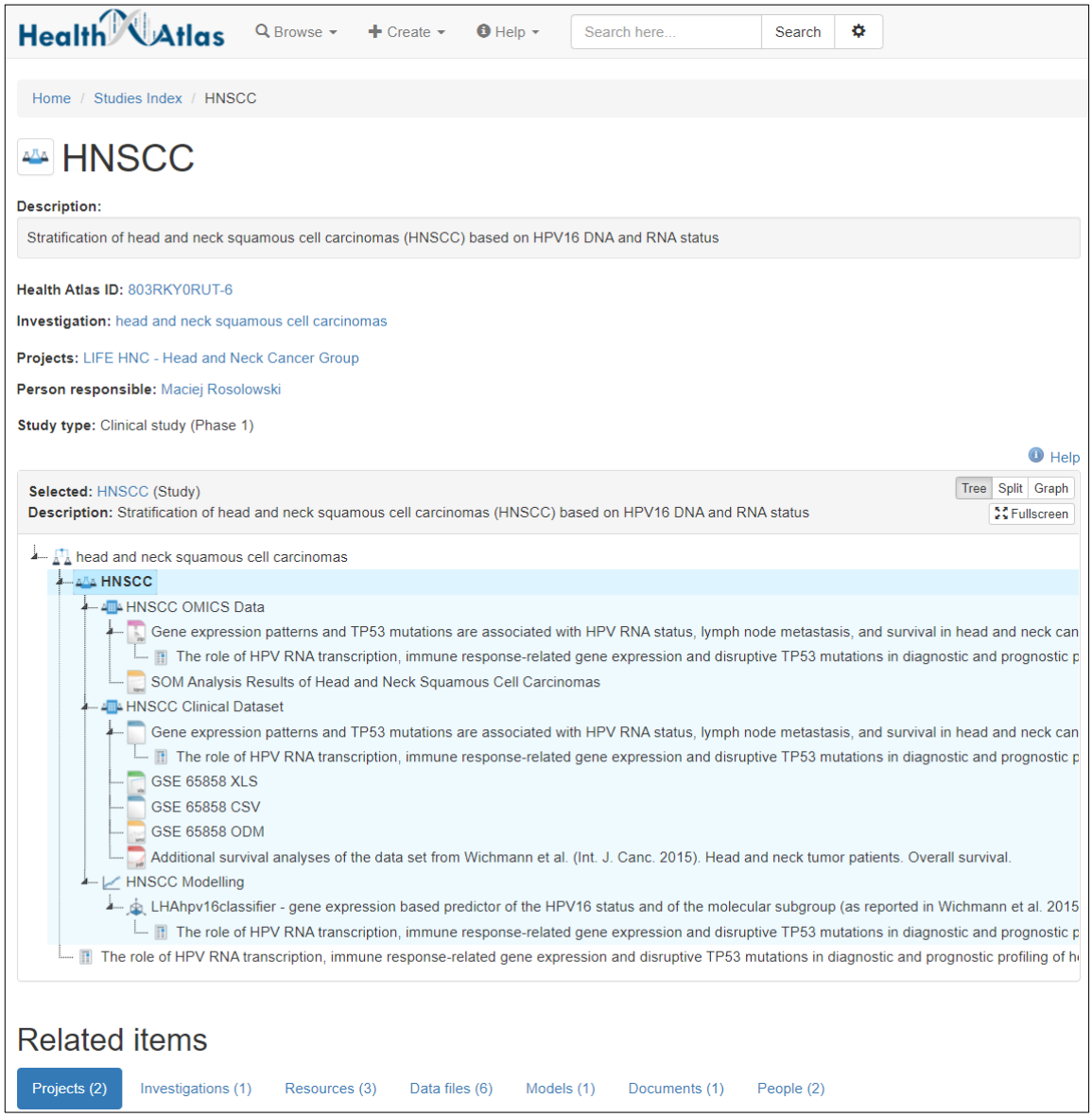

Figure 1 – Screenshot of the Head and Neck Cancer Study (HNSCC) in LHA [https://www.health-atlas.de/lha/803RKY0RUT-6] (cropped for readibility). The figure shows both the Investigation, which is higher-level in a logical sense, and the datasets (SEEK Assay) generated in the study (OMICS, clinical, model) and their binary expressions in different formats. The distributions are freely downloadable, the model is visible only to logged-in users. Descriptive metadata is at the top, associative references at the bottom.

```
<jerm:Experimental_assay rdf:about="http://localhost:3000/assays/1">
      <dcterms:type rdf:resource="http://www.w3.org/ns/dcat#Dataset"/>
      <dcterms:title>HNSCC Clinical Dataset</dcterms:title>
      <jerm:title>HNSCC Clinical Dataset</jerm:title>
      <dcterms:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2021-05-12T16:26:49…
      <dcterms:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2021-05-12T16:36:…
      <dcat:distribution rdf:resource="http://localhost:3000/data_files/1"/>
      <dcat:distribution rdf:resource="http://localhost:3000/data_files/2"/>
      <prov:wasAttributedTo rdf:resource="http://localhost:3000/people/1"/>
      <prov:wasGeneratedBy rdf:resource="http://localhost:3000/projects/2"/>
      ...
```

Figure 2 – Example RDF output code snippet for DCAT-enabled datasets

Currently, the implementation is evaluated in the LHA development system (see Figure 2).

Table 1 – Mapping of entities from SEEK/ISA to DCAT utilizing the HNSCC example

Object	Class in SEEK/ISA	Class in DCAT
The Leipzig Health Atlas (itself)	-	dcat:Catalog
"head and neck squamous cell..."	Investigation	dcat:Resource
"HNSCC"	Study	dcat:Resource
"HNSCC OMICS Data"	Assay	dcat:Dataset
"Gene expression patterns and TP53 …"	Data file	dcat:Distribution
"SOM Analysis Results of Head and..."	Data file	dcat:Distribution
"HNSCC Clinical Dataset"	Assay	dcat:Dataset
"Gene expression patterns and TP53 …"	Data file	dcat:Distribution
"GSE 65858 XLS"	Data file	dcat:Distribution
"GSE 65858 CSV"	Data file	dcat:Distribution
"GSE 65858 ODM"	Data file	dcat:Distribution
"Additional survival analyses of the …"	Document	foaf:Document
"HNSCC Modelling"	Assay	dcat:Dataset
"LHAhpv16classifier"	Model	dcat:Distribution
"The role of HPV RNA transcription …"	Publication	foaf:Document
Maciej Rosolowski (Creator)	People	foaf:Person
René Hänsel (Submitter)	People	foaf:Person
"LIFE HNC"	Project	foaf:Project
"LIFE"	Project	foaf:Project
"Institute for Medical Informatics…"	Institution	foaf:Organization
The i2b2 research data warehouse	-	dcat:DataService

Table 2 – Mapping of classes and methods from SEEK/ISA to DCAT/RDF as used in Figure 2 (excerpt).

Class in SEEK	Method in SEEK	Property in RDF
*	title	RDF::Vocab::DC.title
*	title	JERMVocab.title
*	created_at	RDF::Vocab::DC.issued
*	updated_at	RDF::Vocab::DC.modified
Assay	assets	RDF::Vocab::DCAT.distribution
*	managers	RDF::Vocab::PROV.wasAttributedTo
Assay	projects	RDF::Vocab::PROV.wasGeneratedBy

Discussion

Mapping from SEEK to DCAT was possible for all major entities. As far as the constructs directly from the namespace of DCAT were not sufficient, it was possible to map with constructs from the vocabularies on which DCAT itself is based like Dublin Core, FOAF or W3C PROV[1] and whose use belongs to the best practices. Our implementation is currently only prototypical, but we plan to feed the conceptual design of the DCAT/RDF export back into the core development branch.

Along our complex example scenario, it became clear that we urgently need better metadata along the data sets to achieve unambigous mapping to DCAT. For example, datasets are currently not explicitly related to each other and are not consistently tagged thematically. Datasets need to be characterized in more detail for medical research purposes. Important key data on population, endpoints, number of cases, inclusion and exclusion criteria, documentation forms used, etc. are missing. Appropriate domain ontologies need to be developed for this purpose. The origin of the data and the processing steps are not indicated (data/workflow provenance). Essential statements on the quality of the data contained are missing, especially in the area of secondary use for clinical trials. Furthermore, there are no machine-readable policies for access restrictions to personal data, such as restriction to countries within a certain level of data protection, so that manual inquiries to the operators are necessary. All these extensions and constraints would ideally be implemented in a separate application profile.

Currently the SEEK/ISA data model does not provide all these facts. Even if all the functionality to markup this information were available in the SEEK software, it would still need to be provided by the responsible scientists or data stewards. Based on our experience in the LHA, it can be said that many users only fill in a minimal set of metadata due to time constraints. Here, a better understanding and hence a stronger commitment to minimal metadata sets is needed from journals or funders.

In addition to SEEK, many other digital asset management platforms exist such as CKAN[2], DataVerse[3], and Menoci[4], to name a few. Most of them rely on individual data models. The penetration of DCAT as a common vocabulary with a prospect of a gold standard for FAIR data management also depends on broad tool support. However, other vocabularies targeting the description of digital assets exist besides DCAT, including schema.org[5], which is supported by the major search engine operators, bioschemas.org[6], which was founded in adaption of this name and develops profiled resources for bioinformatics. But standards such as HL7 FHIR[7] are also evolving from the original application field of medical instance data in healthcare to increasingly include the description and characterization of research projects, datasets, or citations.

Conclusions

The research objective of this paper was to improve findability and interoperability of existing data repositories by adding annotation to research data assets with the help of a widely used, consented, domain-relevant, structured, machine-readable vocabulary. DCAT is a stable vocabulary with ongoing development and a large usage base, particularly in the area of government data, but also for indexing services such as Google Dataset Search[8]. The maintaining W3C working group extensively documents requirements, usage scenarios, and implementation evidence. A version 3 is currently already in progress. Mapping to the well established SEEK platform demonstrates the broad

[1] https://www.w3.org/TR/vocab-dcat-2/#references
[2] https://ckan.org/
[3] https://dataverse.org/
[4] https://menoci.io/
[5] https://schema.org/
[6] https://bioschemas.org/
[7] https://www.hl7.org/fhir/
[8] https://datasetsearch.research.google.com/

applicability in the area of research data management in biomedical informatics and medical statistics. A deepening of the semantic relationships with respect to provenance, aspects of data quality and process chains of data lineage is the subject of further work. The use of RDF as a knowledge description language guarantees both sufficient expressive power and a fit with existing services and tools in the field.

DCAT enables a research data center to describe datasets and data services in a repository using a domain-independent, standard vocabulary that facilitates the retrieval, interpretation, and further processing of metadata from its catalog. This can increase the discoverability of datasets and data services. Additionally, it also enables a decentralized approach to publishing data catalogs *across research data centers* and makes *federated searching* for datasets across catalogs in multiple locations using the same query mechanism and structure possible. Exactly this use case is a current challenge in building the distributed national research data infrastructure for Germany within the NFDI4Health project[9].

Acknowledgements

Funding from the German Research Foundation (DFG NMDR grants IN 50/3-2, SA 1009/3-2, WI 1605/10-2 and NFDI4Health DFG grant 442326535) as well as the German Federal Ministry of Education and Research (BMBF Grant 031L0026) is acknowledged.

REFERENCES

[1] M.D. Wilkinson, M. Dumontier, I.J.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, The FAIR Guiding Principles for scientific data management and stewardship. Sci Data **3** (2016), 160018.

[2] K. de Smedt, D. Koureas, and P. Wittenburg, FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications **8** (2020), 21.

[3] Research Data Alliance FAIR Data Maturity Model Working Group, *FAIR Data Maturity Model: specification and guidelines,* Research Data Alliance, 2020.

[4] M. Witt, *2,000 Data Repositories and Science Europe's Framework for Discipline-specific Research Data Management,* DataCite, 2018.

[5] M. Haendel, A. Su, J. McMurry, and Et Al, *Fair-Tlc: Metrics To Assess Value Of Biomedical Digital Repositories: Response To Rfi Not-Od-16-133,* Zenodo, 2016.

[6] A. Perego, D. Browning, R. Albertoni, S. Cox, P. Winstanley, and A. Gonzalez Beltran, *Data Catalog Vocabulary (DCAT) - Version 2,* 2020.

[7] K. Wolstencroft, S. Owen, O. Krebs, Q. Nguyen, N.J. Stanford, M. Golebiewski, A. Weidemann, M. Bittkowski, L. An, D. Shockley, J.L. Snoep, W. Mueller, and C. Goble, SEEK: a systems biology data and model management platform. BMC Syst Biol **9** (2015), 33.

[8] K. Wolstencroft, O. Krebs, J.L. Snoep, N.J. Stanford, F. Bacall, M. Golebiewski, R. Kuzyakiv, Q. Nguyen, S. Owen, S. Soiland-Reyes, J. Straszewski, D.D. van Niekerk, A.R. Williams, L. Malmström, B. Rinn, W. Müller, and C. Goble, FAIRDOMHub: a repository and collaboration environment for sharing systems biology research. Nucleic Acids Res **45** (2017), D404-D407.

[9] C.O. Schmidt, J. Darms, A. Shutsko, M. Löbe, R. Nagrani, B. Lindstädt, M. Golebiewski, S. Koleva, T. Bender, U. Sax, X. Hu, M. Lieser, V. Junker, M. Lehne, A. Zeleke, I. Pigeot, and J. Fluck, Facilitating study and item level browsing for clinical and epidemiological COVID-19 studies. Studies in Health Technology and Informatics (2021), accepted.

[10] M. Parciak, T. Bender, U. Sax, and C.R. Bauer, Applying FAIRness: Redesigning a Biomedical Informatics Research Data Management Pipeline. Methods Inf Med **58** (2019), 229–234.

[11] S.-A. Sansone, P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral-Zettler, K. Begley, T. Booth, L. Bougueleret, G. Burns, B. Chapman, T. Clark, L.-A. Coleman, J. Copeland, S. Das, A. de Daruvar, P. de Matos, I. Dix, S. Edmunds, C.T. Evelo, M.J. Forster, P. Gaudet, J. Gilbert, C. Goble, J.L. Griffin, D. Jacob, J. Kleinjans, L. Harland, K. Haug, H. Hermjakob, S.J. Ho Sui, A. Laederach, S. Liang, S. Marshall, A. McGrath, E. Merrill, D. Reilly, M. Roux, C.E. Shamu, C.A. Shang, C. Steinbeck, A. Trefethen, B. Williams-Jones, K. Wolstencroft, I. Xenarios, and W. Hide, Toward interoperable bioscience data. Nat Genet **44** (2012), 121–126.

[12] Meineke Frank A., Löbe Matthias, and Stäubert Sebastian, Introducing Technical Aspects of Research Data Management in the Leipzig Health Atlas. Studies in Health Technology and Informatics **247** (2018), 426–430.

[13] G. Wichmann, M. Rosolowski, K. Krohn, M. Kreuz, A. Boehm, A. Reiche, U. Scharrer, D. Halama, J. Bertolini, U. Bauer, D. Holzinger, M. Pawlita, J. Hess, C. Engel, D. Hasenclever, M. Scholz, P. Ahnert, H. Kirsten, A. Hemprich, C. Wittekind, O. Herbarth, F. Horn, A. Dietz, and M. Loeffler, The role of HPV RNA transcription, immune response-related gene expression and disruptive TP53 mutations in diagnostic and prognostic profiling of head and neck cancer. Int J Cancer **137** (2015), 2846–2857.

[14] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, and I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc **17** (2010), 124–130.

Address for correspondence

Matthias Löbe, Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE), Universität Leipzig, Härtelstraße 16-18, 04107 Leipzig; Email: matthias.loebe@imise.uni-leipzig.de.

[9] https://www.nfdi4health.de/