Secure distributed machine learning in healthcare: a study on FAIR, compliance and cybersecurity for federated learning

Plug, R.B.F.

Part II

Article 2

Note on Prior Publication

This chapter is equivalent to the following published conference article for which I am the first author:

Plug, R., Liang, Y., Basajja, M., Aktau, A., Jati, P. H. P., Amare, S. Y., Taye, G. T., Mpezamihigo, M., Oladipo, F., van Reisen, M. (2022). *FAIR and GDPR Compliant Population Health Data Generation, Processing and Analytics.* Proceedings of the 13th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences. `urn:nbn:de:0074-3127-1`.

I contributed to the conception, design, implementation, and writing of the study, including the development of the FAIR-compliant data architecture, the data stewardship framework, and the cross-facility federation model. Co-authors supported implementation, coordination, and validation in partner countries.

3

FAIR and GDPR Compliant Population Health Data Generation, Processing and Analytics

Ruduan Plug[1], Yan Liang[1], Mariam Basajja[1], Aliya Aktau[1], Putu Hadi Purnama Jati[2], Samson Yohannes Amare[3], Getu Tadele Taye[3], Mouhamad Mpezamihigo[4], Francisca Oladipo[4], Mirjam van Reisen[1,2]

[1] Leiden University, 2311 EZ Leiden, Netherlands

[2] Tilburg University, 5037 AB Tilburg, Netherlands

[3] Mekelle University, 231 Mekelle, Tigray, Ethiopia

[4] Kampala International University, 20000 Kampala, Uganda

Abstract

Generating and analysing patient data in clinical settings is an inherently sensitive process, requiring collaborative effort between clinicians and informaticians to generate value from these data, while mitigating risks to the data subject. As a result, efforts in utilizing external patient data pose significant challenges. We propose a data-centric framework based on the FAIR principles and GDPR guidelines to enhance data management at the point of care. By using the process of data visiting, a cross-facility method for federated data analytics, we can automate generation of novel aggregate data which was previously not realizable. In two sequential studies we show that these techniques, supported by a data stewardship programme, increase community-wide involvement in data generation, improve transparency and trust, provide direct value and data ownership, and enable regulatory and ethically compliant, cross-national data visiting under curated accessibility patterns for federated analytics.

Keywords: FAIR Data, GDPR, Data Management, Data Stewardship, Clinical Data, Biomedical Ontologies, Data Federation, Data Visiting

3.1 Introduction

The generation and management of clinical Electronic Health Record (EHR) data requires strong safeguards on adherence to regulations, data security and protection of patient privacy and confidentiality [1]. These factors complicate facilitation of regional analytics and data exchange, which is seen as a critical factor in concerns of global and cross-national population health. Various methods have been developed to address concerns of data security and privacy protection [2, 3]. However, these methods tend to be problematic in practical use and lack ontology-based standards for cross-facility interoperability and the versatility to enable adherence to regulations set out by the relevant national Ministry of Health (MoH) and regional legislature.

The study implemented by the Virus Outbreak Data Network (VODAN) Africa investigates the preparation and use of digital patient data in Africa. The African continent is least represented in global health data, and the limitations and challenges on digitisation of health data that lead to biases in globally available data are well-documented [4]. Highly developed nations generate the vast majority of the medical data, and as a result see the most representation and benefit from research, while low-resource and rural areas tend to generate few data and consequently are underrepresented in health research.

Efforts of developed nations to generate digital patient data from remote and impoverished regions with vulnerable populations have often led to extractive practices [5, 6], producing data sets that do not become available to or directly serve the benefit of local populations and their health facilities. The transfer of patient data aggregates from the facility where the data is produced to external research facilities, poses ethical and legal concerns, in terms of the ownership of the data and the link to the point of care [7].

These practices in data generation have led to a lack of trust due to the absence of standards in data ownership [8] and insufficiency of procedural transparency within the generation and use of the data. Lack of capacity of data analytics within facilities compounds the problem of delaying adaptation of localised information systems within clinics that can enable medical data generation and regional clinical data exchange [9], while these localised data management practices at point of care are essential to the development of trustworthy and legally compliant data generation methods [10, 11]. The lack of ownership and meaningful use of the data further undermines the potential acceptance of digitisation of patient data by the patient and other stakeholders. Hence, the quality and completeness of such data can be affected by the obstacles to adoptation of proposed digitisation processes of electronic patient information [7].

Data management methods based on FAIR have been proposed as data localisation strategies to improve the standards for patient data generation and interoperability, and GDPR was utilised as a baseline standard to bridge the gap to governance, which resulted in two studies implemented in Africa, spanning from April 2020 - September 2020 and October 2020 to October 2021 [4].

3.2 FAIR and GDPR Standards

A central standard for regulatory frameworks within this study is encapsulated by GDPR, which forms the basis of the initial trial by conceptualizing the point-of-care as both data processor and data controller [12]. By using these standards, explicit data ownership for the data subject and full control over data are provided at local levels while allowing for usage of these data under informed consent. Within initial conversations with stakeholders across eight African nations, including Tanzania, Uganda, Ethiopia, Somalia, Nigeria, Kenya, Tunisia and Zimbabwe, this baseline was found to provide sufficient

common ground while being flexible to more stringent regulations layered upon GDPR as required by local regulators [13].

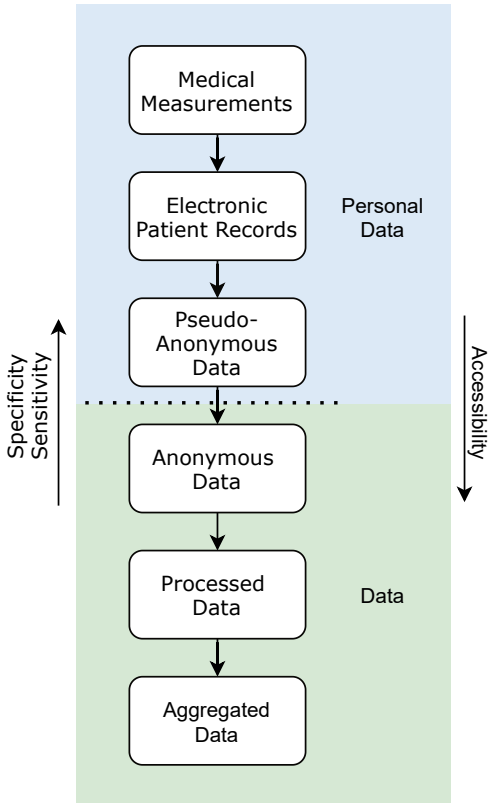

Fig. 3.1.: Levels of Data Processing and Access Control.

An advantage of this approach is that GDPR in itself already provides a legal framework to enable consent-based exchange of processed, anonymised data, requiring an assessment of the relevance of the purpose of the data-collection. However, to enable collaborative use of data such as federated analytics, we have to look towards FAIR and ontology-based metadata to provide transparent, consistent and machine-readable structure to data across different health facilities [14], which can be sourced from HMIS already in use.

By ensuring FAIR compliance at the point of data generation, we provide a set of transparent rules for permissions under which data can be

found and accessed, which is essential in forming trust in management of sensitive data. A six-level system of access is illustrated in Figure 1, in which personal data are not permissible to leave the facility while aggregated, processed and anonymous data may be exchanged through incremental levels of auditing required before clearance is provided [15]. Interoperability is enabled through biomedical ontologies, defined by research communities, providing the semantic links between data which can then be put into practice through metadata templating [16]. The World Health Organisation SMART guidelines recognise the relevance of interoperable digital data use all of these levels, including the importance of the meaningful use of data for quality health access at point of care [17].

3.3 Study Results

To address concerns on security and privacy of patient data, which requires capacities to purposefully address the data production and assignment of responsibilities regarding permission, a data stewardship programme was conceptualised that aims to build a network of local experts on data management and governance [18]. Foundational to a versatile platform of trust and expertise in regard to local and regional circumstances lies the interaction between human domain experts and novel technology, and by bridging this gap, improvements in trust and safety can be attained. Data stewards are primarily trained to handle data management and auditing of data processing directly at the point of care.

Utilizing FAIR, assisted by biomedical ontology services such as NCBO BioPortal [19], has already seen great potential in managing, analysing and reusing biomedical samples across research facilities, for which we show an example in Figure 2. Unique identifiers and data provenance support the documentation of data ownership, while the use of common terminologies and semantics through ontologies ensures

that analytics across facilities is possible. Making such techniques common practice for EHR data makes cross-facility and global analytics of population health data possible without loss of data ownership or extensive post-processing. This is critical for observational research with very limited data such as rare diseases, which impose de-anonymisation risks, or time-sensitive analytics such as measuring incidence of COVID-19 across geographies.

Pathogen: clinical or host-associated sample from Severe acute respiratory syndrome coronavirus 2

Identifiers	BioSample: **SAMN14656635**; Sample name: hCoV-19/USA/WI-179/2020; SRA: SRS6514344	
Organism	Severe acute respiratory syndrome coronavirus 2 Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus; Severe acute respiratory syndrome-related coronavirus	
Package	Pathogen: clinical or host-associated; version 1.0	
Attributes	**strain**	hCoV-19/USA/WI-179/2020
	isolate	Homo sapien
	collected by	Milwaukee Public Health Department
	collection date	2020-03-21
	geographic location	USA: Wisconsin, Milwaukee
	host	Homo sapiens
	host disease	COVID-19
	isolation source	nasal swab
	latitude and longitude	43.042180 N 87.908670 W
	ARTIC barcode identifiers	NB03
BioProject	PRJNA614504 Retrieve all samples from this project	
Submission	UW-Madison, Shelby O'Connor; 2020-04-21	

Accession: **SAMN14656635** ID: 14656635

Fig. 3.2.: An example of rich, ontology-assisted metadata and associated data already being successfully applied and used to enable interoperability and reusability in anonymised pathogen samples isolated from patients [20] (NCBI).

The first study was conducted with universities within Africa, across Uganda, Kenya, Ethiopia, Nigeria, Tunisia and Zimbabwe, in a collaboration of Kampala International University (KIU), Tangaza University, Mekelle University, Addis Ababa University, Ibrahim Badamasi University, University de Sousse, Great Zimbabwe University (GZU), as well as the Leiden University Medical Center (LUMC) [21, 4, 22] in Europe consisting of two core components. The first core component was the sustainable data stewardship programme Training of Trainers (ToT) to train experts in data process curation and data management, based on the FAIR principles under GO TRAIN [23].

The data stewards in turn are also equipped with skills to transfer this expertise to other aspiring data experts, contributing to the UN sustainable development goals [24]. The training program has resulted in 30 trained data stewards whom can produce human and machine readable vocabulary relevant to patient data records [13], from which ontologies can be defined that provide mappings of data to semantics for FAIRification during point-of-care data production. Adhering to the process of building expertise through ToT, the technological architecture was developed and FAIR Data Point (FDP) services were established within clinical settings at medical facilities. The FDPs were implemented using local deployments of DS Wizard [25] to enable FAIR data production, for which data generation was modelled on the WHO SARS-CoV-2 electronic Case Report Forms (eCRF) ontology [26] stored as RDF graph databases.

Following deployment, experiments were performed with local, in-residence data production and subsequent cross-national SPARQL queries using the FAIR data visiting model [27]. The first such clinical query utilizing the findability and accessibility framework of FAIR was held on 29 September 2020 between the FDPs at KIU and LUMC. This study demonstrated the feasibility of data-querying of federated analytics across two continents, involving patient data held in residence, curated and stored in the place where the data was produced.

A successful proof of concept was presented on international regulatory agreements and a clinical implementation of the data ownership preserving framework modelled using the FAIR concepts and GDPR. During this experiment, international cooperation and expertise was developed with focus on findability and accessibility of clinical patient data, findable under well-specified and transparent conditions. The aspects of interoperability and reusability were not operationally implemented during this study and there was only one eCRF as an immutable ontology which limited the flexibility of use.

In direct continuation of the first trial, a second study was conducted to address novel methods to combine ontology-assisted technology and community-expertise in order to enable cross-facility interoperability and ultimately reusability of data [4]. The second study period saw the number of participating nations increase from six to eight including clinics and hospitals from Ethiopia, Kenya, Nigeria, Somalia, Tanzania, Uganda, Tunisia, and Zimbabwe.

Essential to these efforts were retooling and deployment of localised CEDAR [28] instances, which provide an open source platform assisted by BioPortal ontologies to produce, share and curate metadata templates and the data generated from these templates in RDF format. This ensures that data has full providence during production and provides interoperability through the open and transparent definitions of the ontologies. Different templates based on the same ontologies are inherently interoperable on Common Data Elements (CDEs) [29], while data from different ontologies can be matched by similarly utilizing common terms and ontological semantic linkages [30, 31], which match the semantics from one graph structure to another as a translation layer.

Central to the advantages offered by this approach are the engagement of the scientific community, medical facilities, data stewards and legislature, which all have been involved in the design and deployment

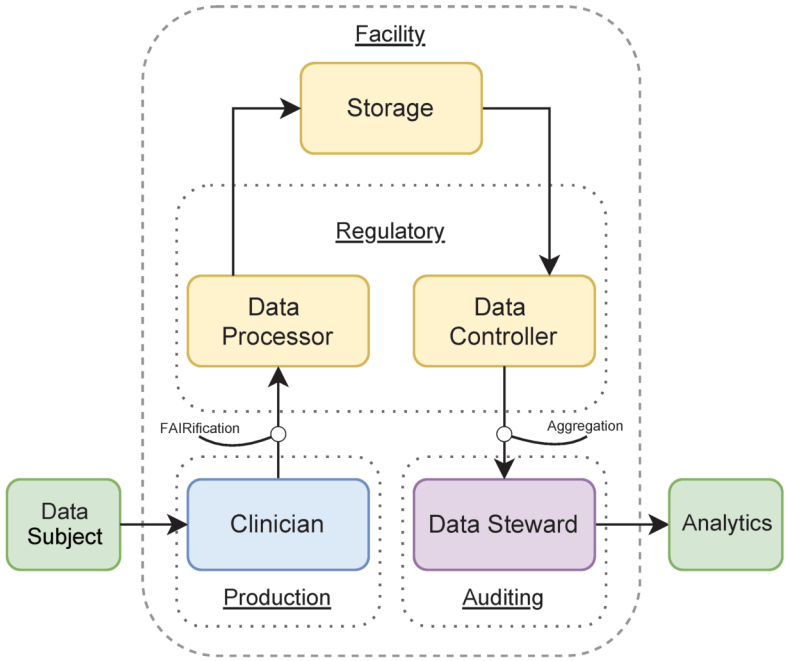

Fig. 3.3.: FAIR and GDPR-based Framework for Data Processing and Analytics.

of this architecture. In addition, broad scale support was received from both the medical community as well as the local MoHs [32]. During the second study, country coordinators have been specified for each country to liaison with local facilities and MoHs, while technical leads form the bridge between country coordinators and the deployment. Data stewards are primarily tasked with guiding and auditing the day-to-day operation of data generation and processing tasks.

The study pioneered a novel, fully FAIR and GDPR compliant, localised health data generation procedure as a distributed network of FDPs that can either function entirely independently or collaborate through data visiting procedures [4]. The resulting minimal viable product resolved the issue of data ownership by fully FAIR local data production being conducted and utilizing expertise from data stewards to conduct audits on data visiting requests, which ensures that all data visiting queries, either to specific facilities or across all indexed FDPs, comply to data

ownership standards and regulations. This is facilitated by means of local data processing, such that the original data never leaves the confinement of the medical facility, towards completely anonymised processed data or aggregates modelled as federated analytics.

The complete procedure of this study is illustrated in Figure 3. This shows the flow of data from the data owner, in this case the data subject, interpreted by local clinicians, processed by data stewards using the FAIR data tooling and then being made available in local storage. Often these data originate from current health information systems such as DHIS2, from which data can also be imported into CEDAR as JSON or RDF formatted data. Upon request for data access using transparent accessibility procedures, under predefined conditions and permission by the data controller, aggregated data can be made available upon clearance of audit by the data steward.

3.4 Conclusion

During this study we have investigated, implemented and deployed a novel FAIR, GDPR compliant data management architecture for curating, repositing and analysing patient health data across health facilities. We have shown that by using the FAIR principles, we can utilise biomedical ontologies to formally structure the data generation process through facility-catered metadata templating, while retaining interoperability among data sources defined by these templates. These formal specifications for interoperability provide an essential component for privacy-oriented federated analytics across health facilities.

With this study we have identified the universal need for the recognition of data ownership and control of patient data in relation to the health facilities where data is produced, and the recognition of data origin and legal rights of the patient as data subject. Data stewardship is proposed as a key instrument in ensuring there is transparency,

community-based trust and accountability for repositing and processing patient data, as well as being instrumental to auditing aggregated analytics performed on these data. This has shown encouraging results with broad support from both health facilities and national MoHs.

In addition, we recognise the importance of the locale of data generation. By keeping full control over the data at the most localised level, we ensure that data are handled in accordance with local regulations and ethical foundations. Based on the support from legislature and research communities, we have found evidence that doing so leads to a higher engagement in data production within previously underserved communities. Broad engagement is essential in reducing data bias and can encourage that aggregated data are being used and analysed in a way that is meaningful within the local context.

By securely repositing data at the most localised level, while exposing curated, rich metadata under FAIR, we enable the possibility for federated data analytics upon individual, controlled authorisation without the risk of exposing the underlying sensitive data. While generating FAIR data can be enabled using a systematic ontology-matching approach, by linking the data generation process to FAIR templates based on domain ontologies, the auditing of data processing and analytical queries still requires significant knowledge and responsibility to comply with ethical standards and local regulations, for which data stewardship forms an essential area of local expertise.

Underlining these findings lies the importance between the relationship of data generation and the in(direct) purpose of such data collection and processing activities. Significant progress in EHR data analytics can be made by improving the processes from the very origin of the data and ensuring that these processes are transparent, well-defined and FAIR, which is in line with the SMART guidelines presented by the World Health Organisation.

References

[1] W Nicholson Price and Ivan Glenn Cohen. „Privacy in the age of medical big data“. In: *Nature Medicine* 25 (2019), pp. 37–43 (cit. on p. 68).

[2] Hao Jin, Yan Luo, Peilong Li, and Jomol P. Mathew. „A Review of Secure and Privacy-Preserving Medical Data Sharing“. In: *IEEE Access* 7 (2019), pp. 61656–61669 (cit. on p. 68).

[3] Ji-Jiang Yang, Jianqiang Li, and Yu Niu. „A hybrid solution for privacy preserving medical data sharing in the cloud environment“. In: *Future Gener. Comput. Syst.* 43-44 (2015), pp. 74–86 (cit. on p. 68).

[4] Mirjam van Reisen, Francisca Onaolapo Oladipo, Mia Stokmans, et al. „Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research“. In: *Advanced Genetics (Hoboken, N.j.)* 2 (2021) (cit. on pp. 68, 69, 73–75).

[5] Florence Femi Odekunle, Raphael Oluseun Odekunle, and Srinivasan Shankar. „Why sub-Saharan Africa lags in electronic health record adoption and possible strategies to increase its adoption in this region“. In: *International Journal of Health Sciences* 11 (2017), pp. 59–64 (cit. on p. 68).

[6] Kathryn M. Chu, Sudha Jayaraman, Patrick Kyamanywa, and Georges Ntakiyiruta. „Building Research Capacity in Africa: Equity and Global Health Collaborations“. In: *PLoS Medicine* 11 (2014) (cit. on p. 68).

[7] Anupam Garrib, Norah Stoops, Andrew Mckenzie, et al. „An evaluation of the District Health Information System in rural South Africa.“ In: *South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde* 98 7 (2008), pp. 549–52 (cit. on pp. 68, 69).

[8] Najia Musolino, J. Lazdiņ, J. Toohey, and C. IJsselmuiden. „COHRED Fairness Index for international collaborative partnerships“. In: *The Lancet* 385 (2015), pp. 1293–1294 (cit. on p. 69).

[9] Mariam Basajja, Marek Suchanek, Getu Tadele Taye, et al. „Proof of Concept and Horizons on deployment of FAIR in the COVID-19 pandemic“. In: *Data Intelligence, Special Issue: Launching an international FAIR data network for COVID data. Forthcoming.* (2021) (cit. on p. 69).

[10] Jeffrey G Shaffer, Seydou Doumbia, Daouda Ndiaye, et al. „Development of a data collection and management system in West Africa: challenges and sustainability“. In: *Infectious Diseases of Poverty* 7 (2018) (cit. on p. 69).

[11] Sundeep Sahay, Arash Rashidian, and Henry Victor Doctor. „Challenges and opportunities of using DHIS2 to strengthen health information systems in the Eastern Mediterranean Region: A regional approach“. In: *The Electronic Journal of Information Systems in Developing Countries* 86 (2020) (cit. on p. 69).

[12] European Parliament & Council. „Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)“. In: *Official Journal of the European Union* 119 (2016), pp. 1–88 (cit. on p. 69).

[13] VODAN-Africa. *About VODAN Africa* (cit. on pp. 70, 73).

[14] Barend Mons, Cameron Neylon, Jan Velterop, et al. „Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud“. In: *Inf. Serv. Use* 37 (2017), pp. 49–56 (cit. on p. 70).

[15] Putu Hadi Purnama Jati, Erik Flikkenschild, Bert Meerman, et al. „Data Access, Control, and Privacy Protection on VODAN Africa Architecture“. In: *Data Intelligence, Special Issue: Launching an international FAIR data network for COVID data. Forthcoming.* (2021) (cit. on p. 71).

[16] Jung ran Park. „Metadata Quality in Digital Repositories: A Survey of the Current State of the Art“. In: *Cataloging & Classification Quarterly* 47 (2009), pp. 213 –228 (cit. on p. 71).

[17] G. Mehl, Ö. Tunçalp, Natschja Ratanaprayul, et al. „WHO SMART guidelines: optimising country-level use of guideline recommendations in the digital age.“ In: *The Lancet. Digital health* (2021) (cit. on p. 71).

[18] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. „The FAIR Guiding Principles for scientific data management and stewardship“. In: *Scientific Data* 3 (2016) (cit. on p. 71).

[19] Patricia L. Whetzel, Natasha Noy, Nigam Haresh Shah, et al. „BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications“. In: *Nucleic Acids Research* 39 (2011), W541–W545 (cit. on p. 71).

[20] Lynn M. Schriml, Maria Chuvochina, Neil Davies, et al. „COVID-19 pandemic reveals the peril of ignoring metadata standards“. In: *Scientific Data* 7 (2020) (cit. on p. 72).

[21] Annika Jacobsen, Ricardo de Miranda Azevedo, Nick S. Juty, et al. „FAIR Principles: Interpretations and Implementation Considerations“. In: *Data Intelligence* 2 (2020), pp. 10–29 (cit. on p. 73).

[22] Mirjam van Reisen, Mia Stokmans, Mariam Basajja, et al. „Towards the Tipping Point for FAIR Implementation“. In: *Data Intelligence* 2 (2020), pp. 264–275 (cit. on p. 73).

[23] Erik Schultes, Albert Mons, Barend Mons, et al. *GO TRAIN Pillar* (cit. on p. 73).

[24] Mirjam van Reisen, Mia Stokmans, Munyaradzi Mawere, et al. „FAIR Practices in Africa“. In: *Data Intelligence* 2 (2020), pp. 246–256 (cit. on p. 73).

[25] Robert Pergl, Rob W.W. Hooft, M. Suchánek, Vojtech Knaisl, and Jan Slifka. „"Data Stewardship Wizard": A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning“. In: *Data Sci. J.* 18 (2019), p. 59 (cit. on p. 73).

[26] Luiz Bonino. *WHO COVID-19 Rapid Version CRF semantic data model* (cit. on p. 73).

[27] Ruduan Plug, Yan Liang, Aliya Aktau, et al. „Terminology on a FAIR-framework for the Virus Outbreak Data Network“. In: *Data Intelligence, Special Issue: Launching an international FAIR data network for COVID data. Forthcoming.* (2021) (cit. on p. 73).

[28] Rafael S Gonçalves, M. O'Connor, M. M. Romero, et al. „The CEDAR Workbench: An Ontology-Assisted Environment for Authoring Metadata that Describe Scientific Experiments“. In: *The semantic Web–ISWC: International Semantic Web Conference proceedings. International Semantic Web Conference* 10588 (2017), pp. 103–110 (cit. on p. 74).

[29] Ching-Heng Lin, Nai-Yuan Wu, and Der-Ming Liou. „A multi-technique approach to bridge electronic case report form design and data standard adoption“. In: *Journal of biomedical informatics* 53 (2015), pp. 49–57 (cit. on p. 74).

[30] Jérôme Euzenat and Pavel Shvaiko. „Ontology Matching“. In: *Springer Berlin Heidelberg*. 2013 (cit. on p. 74).

[31] Pavel Shvaiko and Jérôme Euzenat. „Ontology Matching: State of the Art and Future Challenges“. In: *IEEE Transactions on Knowledge and Data Engineering* 25 (2013), pp. 158–176 (cit. on p. 74).

[32] VODAN-Africa. *VODAN Letters of Support* (cit. on p. 75).