REDbox: a comprehensive semantic framework for data collection and management in tuberculosis research

Vinicius Costa Lima (✉ viniciuslima@usp.br)
Universidade de São Paulo

Filipe Andrade Bernardi
Universidade de São Paulo

Felipe Carvalho Pellison
Universidade de São Paulo

Francisco Barbosa Júnior
Universidade de São Paulo

Márcio Elói Filho
Universidade de São Paulo

Domingos Alves
Universidade de São Paulo

Afrânio Lineu Kritski
Federal University of Rio de Janeiro

Rafael Mello Galliez
Federal University of Rio de Janeiro

Rui Pedro Charters Lopes Rijo
Instituto Politécnico de Leiria

Research Article

Keywords: Tuberculosis, Electronic data capture, Health services research, Data collection, Data management, Medical informatics

Posted Date: November 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1070973/v1

Additional Declarations: No competing interests reported.

Abstract

The outcomes of a clinical research directly depend on the correct definition of the research protocol, the data collection strategy and the data management plan. Furthermore, researchers often need to work within challenging contexts, such as in Tuberculosis services, where human and technological resources for research may be rare. The use of Electronic Data Capture systems, such as REDCap and KoBotoolbox, can help to mitigate such risks and to enable a reliable environment to conduct health research and promote results dissemination and data reusability. The proposed solution was based on needs pinpointed by researchers, considering the lack of an embracing solution to conduct research in low resources environments. The REDbox framework was built to enhance data collection, management and sharing in tuberculosis research, while providing a better user experience. The relevance of this article lies in the innovative approach to support TB research by combining existing technologies and developing support features. When focusing on positive aspects of each tool, it is possible to underpin tuberculosis research by improving data collection, management capability and security. Furthermore, the aggregation of meaning in raw data helps to promote the quality and the availability of research data.

Introduction

Data collection is one of the most crucial moments in all types of research projects, and it could drive a project to success or failure. The lack of quality of data sometimes is noted when the collection phase is over or almost over. To avoid this, in addition to a trained data collector [1], it is essential the use of a reliable data capture system.

Additionally, the success of a clinical research directly depends on the correct definition of the research protocol, the data collection strategy and the data management plan [2]. These elements drive the quality and reliability of the collected data that will be used for analysis of outcomes of a given study.

The adoption of new methods, tools and sources of data have changed the way research is conducted. However, new challenges have arisen, demanding innovative approaches to collect, manage, and publish data. Well-managed data are easier to use and analyze towards the confirmation of a research hypothesis. Also, the reuse of data in further studies is enhanced. In order words, it stimulates more collaboration between researchers and maximizes the investment of funders [3].

The use of an Electronic Data Capture (EDC) system can mitigate the risk of storing potentially sensitive data on paper and help to ensure compliance with medical data privacy, security, and regulations, while improving data quality, management capability and reducing time and costs [4]. An EDC system should be capable to work independent of an operative system or proprietary protocols, and be interoperable, i.e., able to communicate with other systems in a transparent and consistent way [5].

Moreover, in health research, researchers need to work within different contexts. From facilities with high-end devices available to ones with low availability of resources, such as poor – or none – internet connection or even without reliable electrical power. In the case of Tuberculosis (TB), an infectious and neglected disease [6], resources for research may be rare and the costs to use an EDC could be a limitation. These aspects stand out as barriers for collecting data in TB research and, therefore, making data available for further data-driven studies is crucial to underpin the development of new evidence-based decision-making tools.

Integrating information in larger systems is hampered by the heterogeneity of data formats and data structure. Data must be correctly described to be useful [7]. Then, semantic interoperability is a key consideration in information systems design [8]. It is achieved when one system can understand the context and meaning of information provided by another system [9].

Meaning can be added to data by using ontologies or other semantic standards, i.e., well-defined vocabularies which allow precise and machine-readable description of knowledge about a certain domain [10]. It may enable semantic interoperability, allowing systems to make the same interpretation of the data in terms of its formal definition [11]. In this sense, data can be shared in an accurate and reliable way to enhance communication among computerized systems. This capability is especially desirable in health information systems (HIS) due to the heterogeneity of the medical language and health related concepts [9].

Ontologies are important in semantic alignment for data integration, information exchange, and semantic interoperability [12]. An ontology is composed of several properties and each one describes a specific piece of data in the domain being represented [13]. Besides ontologies, simple standards such as the Humanitarian Exchange Language (HXL) help to speed up data processing and create interoperability across data sources. HXL is a project by the United Nations Office for the Coordination of Humanitarian Affair for the coordination of disaster response with semantic web technologies. It uses simple marking through hashtags and its goal is to contribute to the automatization of processes to improve the information flow for decision makers [14].

In the case of health research, semantic annotation can help describe the data that is being collected. It can be useful to later extract and link different research datasets described by the same vocabulary. Usually, each study counts with several collection instruments, totalizing hundreds of fields to be filled during the research progress. Manual annotation is always a choice, but automated approaches for semantic annotation is an extremely important task [15].

Objectives

Clinical trials and studies have increasingly started using EDC systems to conduct a range of analysis [16]. In this sense, and considering that Brazil is part of the top 30 high TB burden countries [17], this work aims to present REDbox, a comprehensive framework based on REDCap [18] and KoBoToolbox [19] systems to enhance research data collection and management in low resources fields, such as TB services, while providing a better user experience.

Additionally, REDbox is intended to promote semantic interoperability of TB research data. Therefore, relying in ontologies and HXL to perform semantic annotations, the objective is to automate the design of an instrument based on a given ontology, and the generation of ontologies derived from instruments' schema, as well as to increase the availability of data for further data-driven TB research.

Methods

In this research, no clinical data and private or public databases were used.

The lack of an embracing solution to conduct research in low resources environments, such as TB reference services in Brazil, led to the conceptualization of the approach proposed in this work. None was found in the literature and after rounds of discussions with researchers, overcoming existing technological barriers in TB services was defined as the main challenge to be faced, such as poor internet connection, unavailability of devices, and staff with low training in digital tools. For a validation phase, REDbox is currently in use in five cross-institutional TB research projects in Brazil. Also, it is demonstrated how the use of semantics can promote data reusability and interoperability of research data.

Therefore, the following research questions were defined:

- "Is it possible to deliver a tool for research data collection and management to be used in low resources environments, such as in tuberculosis services?"
- "How to promote data interoperability to increase availability of TB data for researchers?"

The solution is relevant because it may:

1. Improve the collection and analysis of research data during the whole study period;
2. Facilitate the management of research events and data;
3. Increase the user experience by combining positive aspects of existing solutions;
4. Increase security of research data;
5. Remove technological barriers by delivering an approach that works on any device and without internet connection;
6. Remove cultural barriers, such as the lack of confidence of researchers to drop paper-based methods;
7. Promote semantic interoperability of collected data for data reuse and record linkage.

REDCap and KoBoToolbox

REDCap is a web-based metadata-driven software built in 2004 by a team at Vanderbilt University to enable classical and translational clinical research, basic science research and general surveys, providing researchers with a tool for the design and development of electronic data capture tools [18][20]. REDCap is a free software, but it is not considered open-source. A license is required and it can be installed and managed by a small IT team [21].

Developed by the Harvard Humanitarian Initiative, KoBoToolbox is a free, open-source suite of tools for data collection and basic analysis. It was initially built for use in challenging environments in developing countries [19]. KoBoToolbox is powered by Enketo open-source project [22] and offers online and offline forms availability to be used in any modern browser, thanks to HTML5 features. The software relies on the XLSForm standard, which simplifies the authoring of forms in spreadsheets in a human readable format [23]. A visual and intuitive form builder is available, or forms can be imported as XLS files.

The REDCap system is widely used by the scientific community to collect and manage research data, allowing researchers to conduct their studies by themselves. However, the software may present some usability issues during data collection, such as a polluted graphical interface, gradual performance degradation, and the lack of offline operation without depending on a mobile application.

Although it presents more basic functionalities, the KoBoToolbox delivers modern styles and allows users to work offline directly from the web browser. Therefore, the software may be an important allied to mitigate the usability issues of the REDCap.

Data annotation for semantic interoperability

To better represent collected data, fields in research forms can be annotated with semantic vocabularies. REDCap offers the possibility to include annotations for each field, which will not be displayed on the form or survey, but will be available to the designer and in data exports to help understand the data [20]. This annotation can be a property of an ontology or an HXL hashtag, depending on the user’s preference.

KoBoToolbox natively supports the use of HXL. When authoring an XLSForm, the user must simply insert one extra column in the spreadsheet and fill it with HXL hashtags identifying the type of information in each column. The form builder also provides an intuitive way to relate a hashtag with a instruments’ field.

Results

The framework was developed using PHP v7.4 scripting language [24] is composed of five modules, as follows: i) a metadata database and an Admin System; ii) a Form Converter; iii) an ETL (extract-transform-load) Processor; iv) a Data Quality Module; v) and the Ontology Services. Figure 1 shows the REDbox framework overview.

The metadata database and the Admin System

The web-based Admin System was developed in C# [25] and JavaScript [26] programming language to easily manage the mandatory metadata through create, read, update, and delete (CRUD) operations. Figure 2 presents the relational database model.

In general, first an entry to a REDCap project must be created (table *redcap_project*), including the Application Programing Interface (API) parameters and, then, each project’s instrument must be registered (table *redcap_forms*). The table *form_metadata* stores semantic mapping for instrument’s fields.

Additionally, the following tables are used by the Data Quality Module: *redcap_validation_types*, *redcap_validation_rules*, *redcap_validation_issues*, *redcap_visits*, *redcap_visits_config*, *redcap_alerts*, *redcap_alerts_log*.

The Form Converter

Considering that instruments are built using distinct standards in each software, a converter is desired, so the designer does not have to create the same form twice. Forms in REDCap can be automatically created through derivation from ontologies, or the conversion of a form designed in XLSForm standard.

To initiate the process, the user must upload the spreadsheet (.xls) or the ontology (.owl) file, fill the form name, and choose between generating a .zip file, to manually upload it into REDCap, or automatically importing the form through the API. In the second option, the API Token and URL must be provided. Figure 3 shows the user interface of the converter.

Deriving from ontologies. Each property of a given ontology can be converted to fields in forms. The name and type of a field is obtained from the name of the property and the associated type (text is the default type). Minimum and maximum values defined as restrictions on properties are also converted.

Converting from XLSForms. The converter supports all common field types, such as: text, date, date and time, time, integer, decimal, calculation, single selection, multiple selection, files and notes. These types of fields will be converted as they are, including the variable name and values assigned to options in single and multiple selections, so instruments will have matching structure on both systems. Skip logic defined on KoBoToolbox is translated to REDCap branching logic, as well validations rules.

In the designing process, there is a particularity related to multiple selection questions (checkboxes). This type of question needs to have the field's name starting with '*checkbox_*'. This is needed to ensure a correct identification of a multiple selection question structure during data transfer from KoBoToolbox to REDCap.

Before starting the conversion process, the naming convention will be pre-checked by the converter module. If any inconsistency is detected, the conversion will fail, and the user will be informed with the detected error.

The ETL Processor

After converting the instrument and transmitting it to REDCap, KoBoToolbox native REST Services must be enabled in the form settings to instantly submit collected data to the ETL processor through a POST request. The processor URL and basic HTTP authentication credentials must be provided.

The processor receives the data collected in KoBoToolbox as a JSON object, which is parsed to remove unnecessary elements that are not related to the data of interest. After verifying authentication credentials, the metadata is queried to obtain the URL and the token of the REDCap API (table *redcap_projects*) and to verify if it is the first form in the project (table *redcap_forms*). If it is, a request is sent to REDCap API to generate a new record ID, which means that it is a new participant in a research project. Otherwise, the record ID will be searched in the log of collected data, based on the participant identifier. Then, a request is sent to the REDCap API to import the data.

After successfully saving the data, additional steps may take place depending on the settings defined for the instrument, such as: sending of e-mail notifications (both for the respondent and the research team), verification of duplicity of records, and the instant lock of the saved record (to avoid changes in the data). These are useful features that may facilitate the management of research data.

Once the data is in the REDCap database, changes in records are monitored through the Data Entry Trigger module, which can detect any changes. When it occurs, the processor exports the edited data from REDCap and logs it into the relational database.

Data Quality Module

Data management is a continuous process and represents a critical phase in clinical research, due to its importance to the generation of high-quality and reliable data for statistical analysis, which should meet the protocol-specified parameters and comply with the research protocol requirements [27].

It is crucial that the management activities occur in parallel with the data collection. The data manager usually carries out a data validation process, which includes the verification of the consistency, completeness and accuracy of collected data. That way, it is expected to avoid missing data and an increase in quality.

In health research, most data are acquired during participant’s visits. Therefore, keeping track of the schedule of visits and their status (carried out, not carried out, pending) are essential for not missing any milestone.

However, all of these tasks are time consuming, because they demand a careful inspection of a significant amount of data. The REDCap software natively offers useful tools to help data managers and researchers, such as the Resolution Workflow and the Scheduling features, which allows the opening of queries to request the verification of the collected data and assists in the scheduling of expected visits for participants during the study (although it requires a manual setup for each participant), respectively.

The Data Quality Module is composed of three submodules that can complement the functionalities offered by REDCap, focusing on the reduction of the workload for data managers and researchers.

First, there is an automatic rule-based validation procedure that goes through each field in all instruments searching for any inconsistency. Rules must be pre-defined as metadata and they represent the format or range of values expected for a given field. The procedure runs several times a day to check, at the same time, for new issues and to verify the resolution of previously identified ones. When an issue is detected, a query is opened in the Resolution Workflow (in REDCap) and the data collector is alerted by e-mail. Figure 4 presents the dashboard with an overview of all issues detected in a REDCap project.

Additionally, a panel was developed to provide a quick visualization of all upcoming participants' visits. Each row in the panel is a participant and each column a visit. The color of cells represents the status of a visit (green: carried out; red: not carried out; yellow: pending/waiting for the participant). Dates are calculated based on a reference date field (e.g., the day of an intervention or inclusion in the study) and in the days offset for each event. This information is also stored as metadata. The panel is created in real-time with online data extracted from the REDCap database, saving time of researchers that usually create their own panel using spreadsheets. Figure 5 shows the panel for a study with 21 visits (project IV in Table 2).

Finally, an alert system was designed to periodically send notifications to the research centers regarding pending data collection based on the scheduled events of each study. Through reminders, the system helps researchers to keep the participants data up-to-date according to the formal protocol and, therefore, avoiding critical violations. The notifications are sent by e-mail or SMS messages to the recipients list stored as metadata.

Ontology Service

The solution offers a service that provides practical tools to enhance the use of ontologies in the system and allow the continuous integration of different data sources, able to adapt to the evolution of ontologies and ensure availability and avoid data loss.

As previously stated, the form converter is able to derive an instrument from an ontology. In a similar way, this service enables the creation of an ontology based on an instrument. This feature relies on an external application, namely the D2R Server [28, 29]. The D2R is a tool that converts relational contents in semantic formats, allowing a quick conversion between these formats by automatically creating ontologies based on the schema of the content.

Relying on this feature, REDbox can define an ontology from a data collection instrument. To achieve this, a temporary table is created on a relational database, where each column represents a field in the instrument. Then, the D2R generates and publishes an ontology using the table structure, i.e., converting columns to properties, which can be later customized. Table 1 presents an example of an ontology generated from an instrument containing patient's treatment data.

Table 1
Instrument and ontology correspondence

Instrument		**Ontology**	
TB treatment			
Field	**Type**	**Property**	**Range**
Start date	textbox with date validation	http://vocab.redbox.technology/vocab/treatment/start_date	Literal (date)
TB clinical form	Multiple choice with single answer	http://vocab.redbox.technology /vocab/treatment/clinical_form	Literal
Discharge date	textbox with date validation	http://vocab.redbox.technology /vocab/treatment/discharg_date	Literal (date)

The Ontology Service guarantees the semantic interoperability between the applications and formularies that use different versions of the same ontology or even between different ontologies by maintaining the history of changes and mapping the concepts from one ontology version to another. There are a few features that this piece of software contemplates: upload of a file containing the source term of one ontology version and correspondent target one in the new ontology version; upload annotated files with one ontology version and convert them to an older/newer version of the same ontology; or upload a marked-up file with an ontology and convert it to a file of correlated ontology that was previously aligned/mapped.

Validation

The validation of the proposed solution is performed by its use in several cross-institutional research projects related to TB in Brazil, namely: I) Longitudinal Study of the Impact of Social Support on Tuberculosis Indicators - ELISIOS; II) Validation of the Line Probe Assay's performance as a rapid diagnosis method for drug-resistant tuberculosis in reference centers in Brazil; III) Validation of Recombinant PPD in the Diagnosis of Tuberculosis Infection; IV) ProBCG - Use of the Bacillus Calmette–Guérin (BCG) vaccine as prevention of COVID-19 in health professionals. Table 2 shows the characteristics of each project that are currently using the framework; and V) Multicenter prospective clinical trial to assess the diagnostic accuracy of the Truenat method for routine use.

Table 2
Characteristics of each project that are currently using REDbox

Project	No. of research centers	No. of instruments	No. of fields	Expected no. of records
I	3	4	175	2500
II	14	14	679	3800
III	7	9	180	1020
IV	3	24	528	1000
V	5	14	357	500
Total	32	65	1919	8820

It is possible to note that there is a significant number of instruments and fields on each project. That is to say that the form converter module is crucial in this scenario, where each form needs to be designed only once in KoBoToolbox and, then, converted to the REDCap format. The expected number of records is also significant, which may demand the use of easy- to-use and offline tools.

So far, the main benefits reported by end users of research centers relate to the ability to collect data in interviews with patients in scenarios with unstable internet connection, receiving personalized alerts based on events, and the possibility to quickly visualize the expected visits of each participant during the study.

Discussion

The relevance of this article lies in the innovative approach to support TB research during collection and management phases, which is often carried out in contexts with few human and technological resources. These phases can be improved through the REDbox framework, which offers useful tools and a better user experience based on the integration of the REDCap and KoBoToolbox EDC systems and the use of semantics.

The main motivation for this work was to allow health research to be carried out in TB services, where, in general, technological resources are scarce and precarious. Thus, the proposed solution allows the search to facilitate the collection and management of research data. Despite being based on the TB context, the framework can be applied in any area with the same demands.

Although REDCap presents a better approach to the whole research life cycle, some usability concerns and offline availability could be a significant drawback. REDCap has a mobile app that enables offline data collection, but it may not be enough, due to the dependency of smartphones and/or tables availability in research centers, the poor usability provided, and the non-compatibility of some advanced features [30]. Also, mobile devices in digital data collection projects are frequently not owned by the people entering the data, which can be considered a risk to be managed [31].

On the other hand, KoBoToolbox natively works directly from a mobile browser, without any additional software. Due to the use of HTML5 features, KoBoToolbox provides a better user experience through modern forms styles and a way to work offline if needed, without the use of any additional application, such as mobile apps.

Semantics. The semantic annotation can underpin the exchange, use and integration of data from different sources thanks to the aggregation of meaning in raw data. In other words, data becomes machine understandable and can be interpreted by distinct systems.

In the research project IV, as shown in Table 2, a semantic integration has been performed using data collected in the research's instruments and in HIS from the Brazilian Ministry of Health. In this case, demographic and vaccination information were integrated and compared to keep data up to date and increase completeness in the research dataset.

APIs. APIs enable interoperability and data integration between software components and the development of extensions of existing systems.

Regarding REDCap, the API is well-documented, and several endpoints are available, which basically allows managing a whole project programmatically. In this work, some endpoints were used, specifically to: i) import and export data; ii) import files; iii) generate unique identifiers (record id); iv) import metadata (instruments, fields); and v) export metadata.

In KoBoToolbox, the API is not adequately documented. However, a feature is available to instantly send collected data to an external server in JSON standard. This is very useful when using the system only for data collection, which is the intention of this work, and because it eliminates the need of developing a client to extract data.

Data safety. In general, data is stored in three distinct logical units, such as KoBoToolbox database, REDCap database, and in the relational database. Only data stored in REDCap is intended for analysis, but in case of any failure, data can be easily restored. Finally, the whole process is transparent to the final user, which can focus only on data collection, management, and analysis.

Limitations. In the form converter, the designer must pay attention in the following aspects:

1. i) *need of using a variable naming convention for multiple selection fields (checkboxes).* Using a naming convention for variables of multiple selection fields is crucial. Otherwise, data transferring may fail.
2. ii) *calculated fields.* When using calculated fields, KoBoToolbox does not allow setting up a label for this kind of field, unlike REDCap. As a workaround, the designer can use the “Guidance Hint” option, which will be transformed into a label when converted to REDCap format. However, this is not mandatory since REDCap accepts blank labels in calculated fields.

A drawback of using the REDbox framework include the need to define several configuration parameters in the metadata database to the proper functioning of the system and the effective integration of REDCap and KoBoToolbox. It may represent a workload in the initial phase of the research project (the setup must be carried out before starting the data collection), which varies according to the modules that will be used. The Admin System and user's manual seek to make this task easier, but some technical knowledge may be necessary for a correct configuration.

Conclusions

This work has presented REDbox, a comprehensive framework for integrated data collection and management in tuberculosis research. The use of REDCap and KoBoToolbox together has allowed the combination of the advantages of each one, in a transparent way, helping researchers to manage and maintain data while increasing satisfaction of final users that are responsible for collecting data in the field. Furthermore, the Data Quality module intends to speed up and enhance data management by reducing the workload of time-consuming and delicate tasks.

Supporting the semantic integration of data is also another important contribution of this work. The addition of meaning in raw data and the possibility to follow the evolution of ontologies through versioning are crucial to promote the quality and the availability of research data over time.

Finally, although the solution was motivated by the TB scenario, it is applicable in other health fields.

Declarations

Acknowledgements

The authors would like to thank the Brazilian Network for Tuberculosis Research (REDE-TB) for providing access to the licensed instances of REDCap and KoBoToolbox used in this study.

This work was partially supported by: the Brazilian National Council for Scientific and Technological Development (CNPq) - grant number 440758/2018-1, coordinated by author A.L.K; the Brazilian Ministry of Health – grant numbers SVS/MS/22410-2019 and SVS/MS/22411-2019 - coordinated by author A.L.K; the São Paulo Research Foundation (FAPESP) - grant number 2020/01975-9, coordinated by author D.A.

Authors' contributions

CRediT (Contributor Roles Taxonomy) author statement

V.C.L.: Conceptualization, Methodology, Software, Writing - Original Draft, Writing - Review & Editing, Validation. F.A.B.: Conceptualization, Validation, Writing - Original Draft; F.C.P.: Validation, Writing - Original Draft; F.B.J.: Methodology, Writing - Review & Editing; M.E.C.F.: Software; D.A.: Project administration, Funding acquisition, Resources, Writing - Review & Editing. A.L.K.: Funding acquisition, Project administration. R.M.G.: Methodology. R.P.C.L.R.: Conceptualization, Methodology, Supervision, Writing - Review & Editing. All authors reviewed the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Law, E. L. C., Roto, V., Hassenzahl, M., Vermeeren, A. P. O. S. & Kort, J. Understanding, scoping and defining user experience: A survey approach. in *Conference on Human Factors in Computing Systems - Proceedings* 719–728 (2009). doi:10.1145/1518701.1518813
2. Lynch, C. Big data: How do your data grow?, **455**, 28–29 (2008).
3. Strasser, C. *Research Data Management: A Primer*. http://www.niso.org/publications/primer-research-data-management (2015)
4. Bart, T. Comparison of electronic data capture with paper data collection: Is there really an advantage? *Bus. Brief*, **30**, 1–4 (2003).
5. Iroju, O., Soriyan, A., Gambo, I. & Olaleke, J. Interoperability in Healthcare: Benefits, Challenges and Resolutions. *Int. J. Innov. Appl. Stud*, **3**, 262–270 (2013).
6. de Oliveira, R. G. Meanings of neglected diseases in the global health agenda: The place of populations and territories. *Cienc. e Saude Coletiva*, **23**, 2291–2302 (2018).
7. Konopka, B. M. Biomedical ontologies - A review. *Biocybern. Biomed. Eng*, **35**, 75–86 (2015).
8. Davies, J., Welch, J., Milward, D. & Harris, S. A formal, scalable approach to semantic interoperability. *Sci. Comput. Program*, **192**, 102426 (2020).
9. Garde, S., Knaup, P., Hovenga, E. J. S. & Heard, S. Towards semantic interoperability for electronic health records: Domain knowledge governance for openEHR archetypes. *Methods Inf. Med*, **46**, 332–343 (2007).
10. Chandrasekaran, B., Josephson, J. R. & Benjamins, V. R. What are ontologies, and why do we need them? *IEEE Intell. Syst*, **14**, 20–26 (1999).
11. Wang, H. Q., Li, J. S., Zhang, Y. F., Suzuki, M. & Araki, K. Creating personalised clinical pathways by semantic interoperability with electronic health records. *Artif. Intell. Med*, **58**, 81–89 (2013).
12. Min, H. *et al.* Towards a standard ontology metadata model. in CEUR Workshop Proceedings vol. 1747(2016)
13. Jepsen, T. C. Just what Is an ontology, anyway? *IT Prof*, **11**, 22–27 (2009).
14. Keßler, C. & Hendrix, C. The Humanitarian eXchange Language: Coordinating disaster response with semantic web technologies. *Semant. Web*, **6**, 5–21 (2015).
15. Singhal, A. & Srivastava, J. Generating semantic annotations for research datasets. *ACM Int. Conf. Proceeding Ser.* (2014) doi:10.1145/2611040.2611056

16. Metke-Jimenez, A., Hansen, D. & FHIRCap Transforming REDCap forms into FHIR resources. in *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* vol. 2019 54–63 (2019)
17. WHO. *Global Tuberculosis Report 2020*. (2020)
18. Harris, P. A. *et al.* Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform*, **42**, 377–381 (2009).
19. Harvard, H. I. KoBoToolbox Website. https://www.kobotoolbox.org/ (2018)
20. Wright, A. & REDCap A Tool for the Electronic Capture of Research Data. *J. Electron. Resour. Med. Libr*, **13**, 197–201 (2016).
21. Klipin, M., Mare, I., Hazelhurst, S. & Kramer, B. The process of installing REDCap, a web based database supporting biomedical research: the first year. *Appl. Clin. Inform*, **5**, 916–929 (2014).
22. Laboratory, S. E. C. U. Enketo - Smart paper. https://enketo.org (2012)
23. Sustainable, E. & Laboratory, C. U. & ODK Team, U. of W. XLSForm documentation. https://xlsform.org/en/
24. Lerdorf, R. P. H. P. Hypertext Preprocessor. https://www.php.net (1994)
25. Microsoft, C. C# documentation. https://docs.microsoft.com/en-us/dotnet/csharp/ (2000)
26. Eich, B. & Community, J. JavaScript Programming Language. https://www.javascript.com (1995)
27. Krishnankutty, B., Bellary, S., Moodahadu, L. S. & R, N. K. B. & Data management in clinical research: An overview. *Indian J. Pharmacol*, **44**, 168–173 (2012).
28. Cyganiak, R. & Bizer, C. D2R Server: A Semantic Web Front-end to Existing Relational Databases.*XML Tage*2–4(2006)
29. Bizer, C. & Cyganiak, R. D2R Server - Publishing Relational Databases on the Semantic Web. *5th Int. Semant. Web Conf.* 26 (2006)
30. Tomko, R. L. *et al.* Using REDCap for ambulatory assessment: Implementation in a clinical trial for smoking cessation to augment in-person data collection. *Am. J. Drug Alcohol Abuse*, **00**, 1–16 (2018).
31. Cobb, C. *et al.* Computer security for data collection technologies. *Dev. Eng*, **3**, 1–11 (2018).

Figures

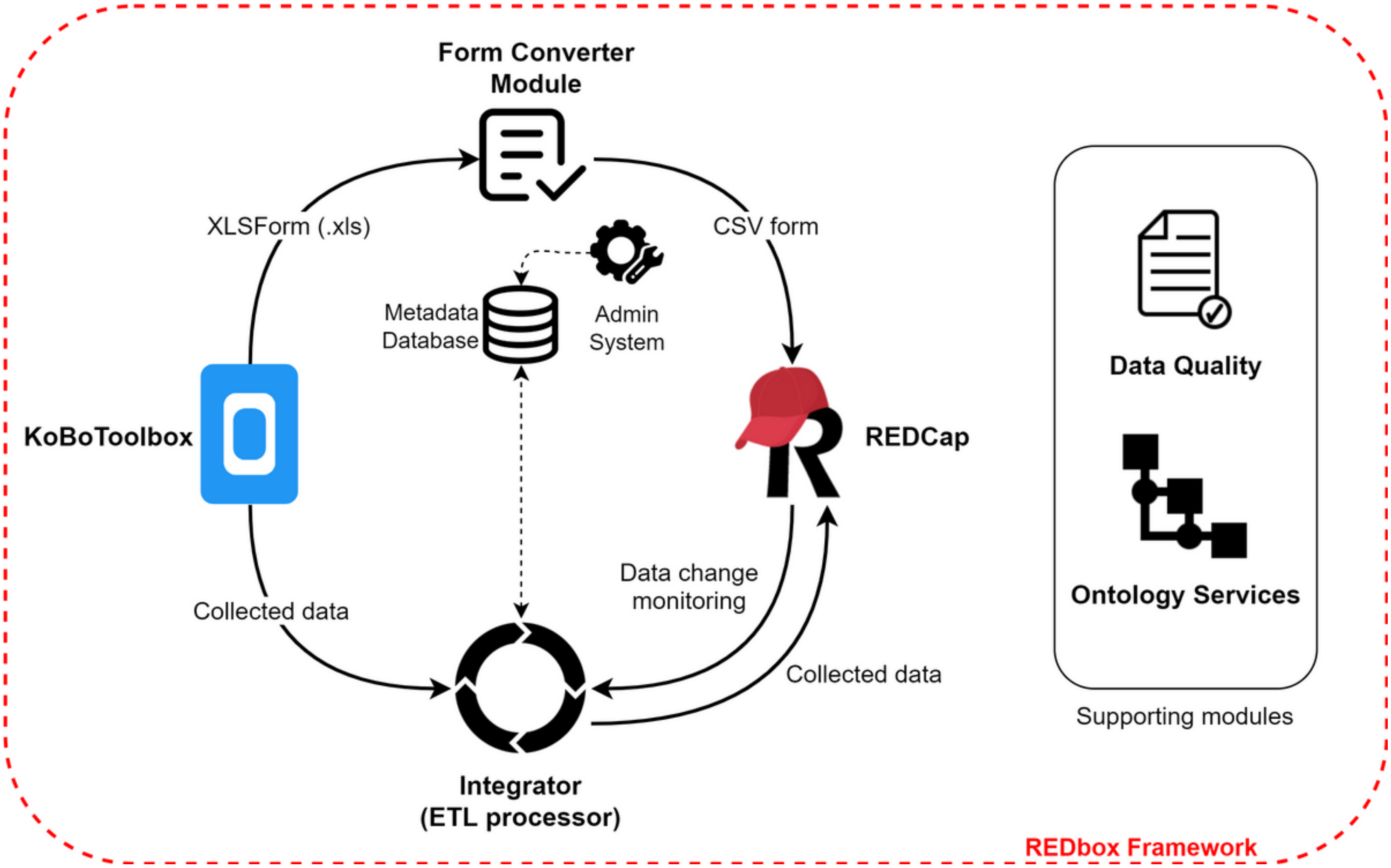


Figure 1

REDbox framework overview

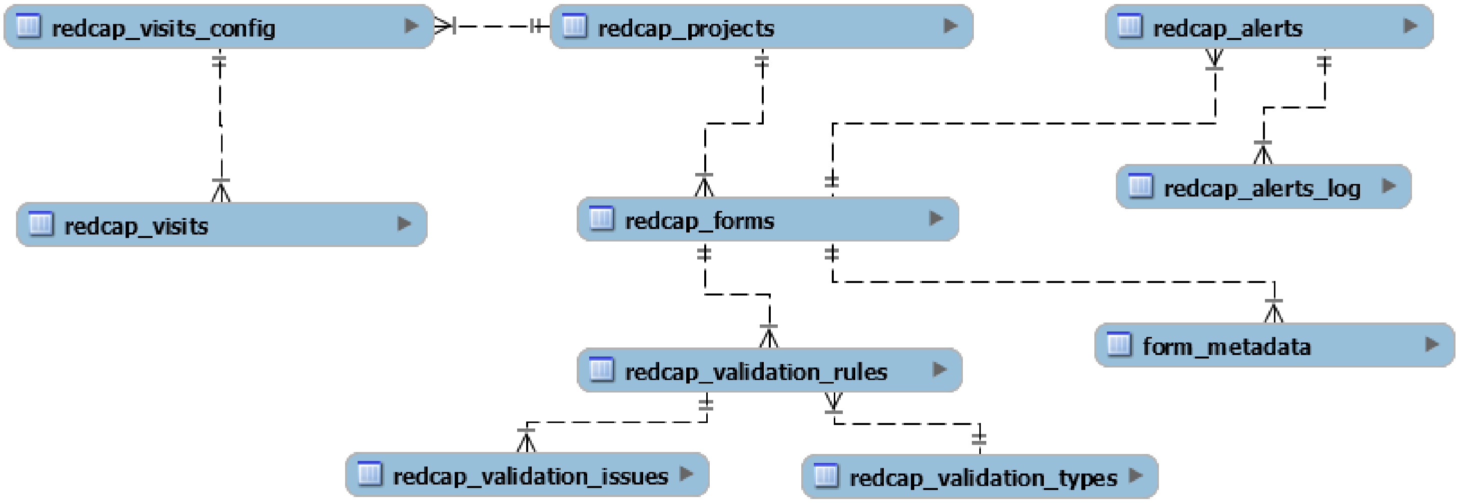
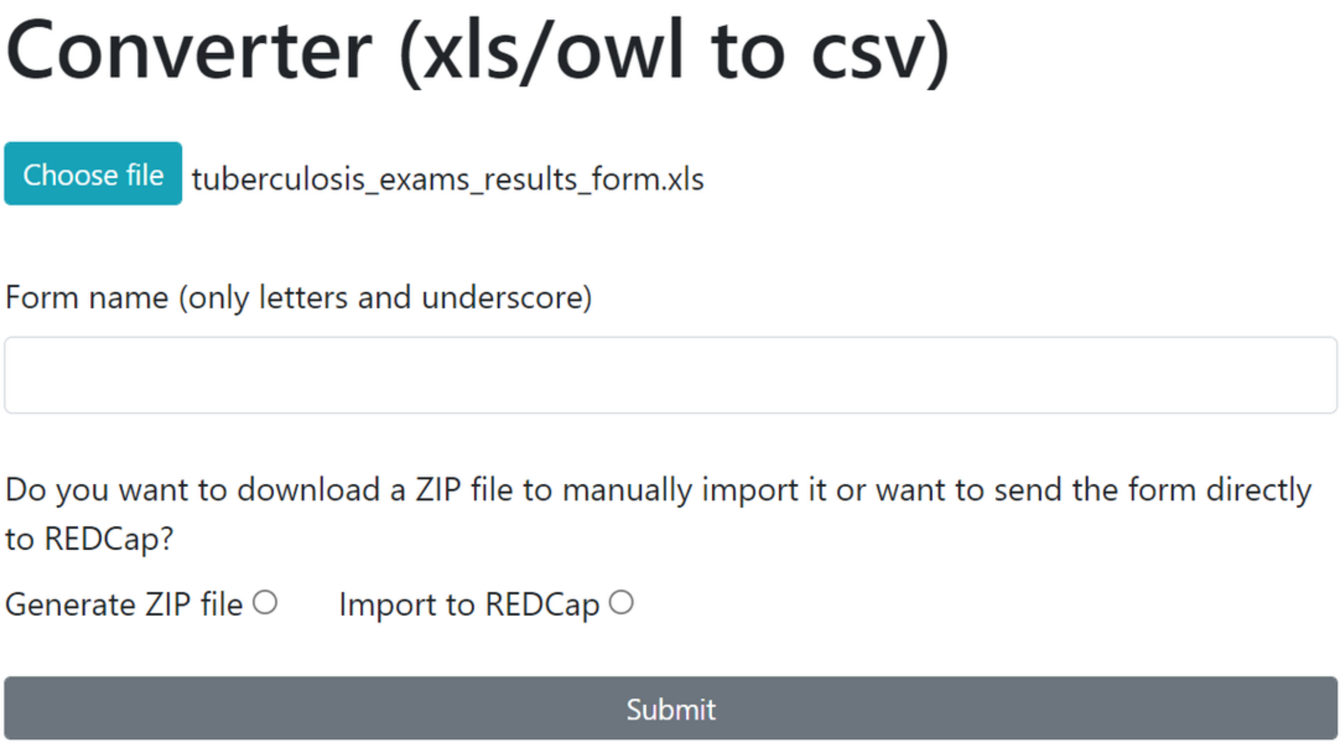


Figure 2

Relational database model



Converter (xls/owl to csv)

tuberculosis_exams_results_form.xls

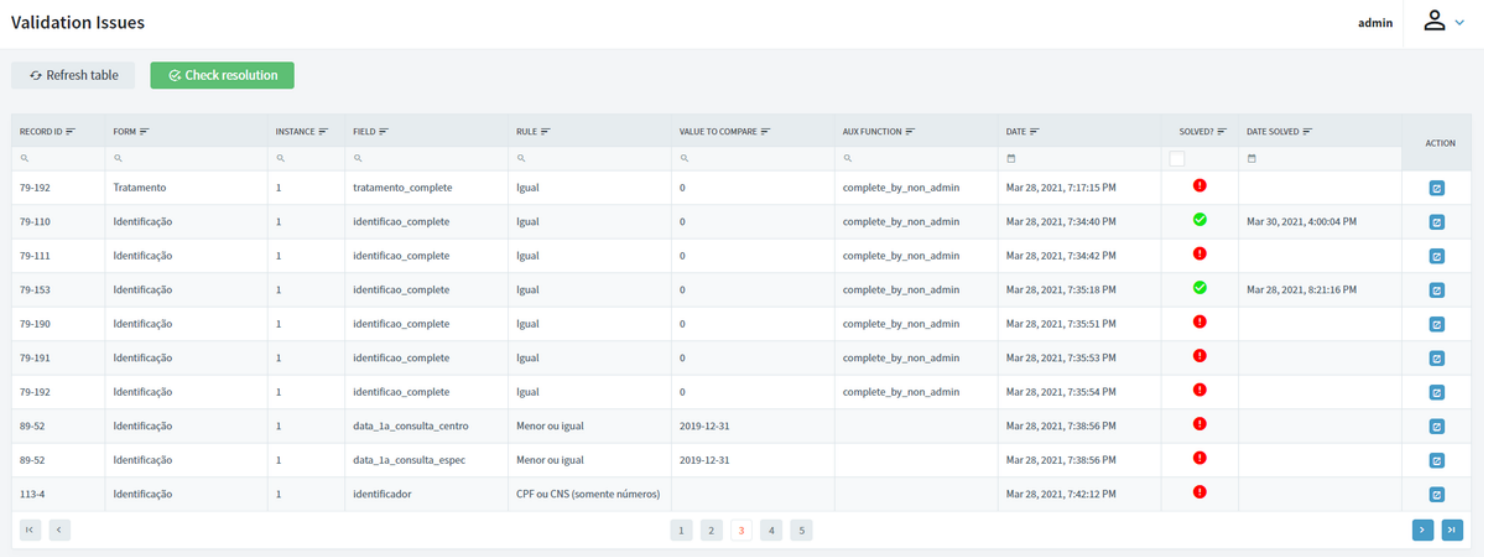
Form name (only letters and underscore)

Do you want to download a ZIP file to manually import it or want to send the form directly to REDCap?

Generate ZIP file ○ Import to REDCap ○

Figure 3

Converter module - user interface

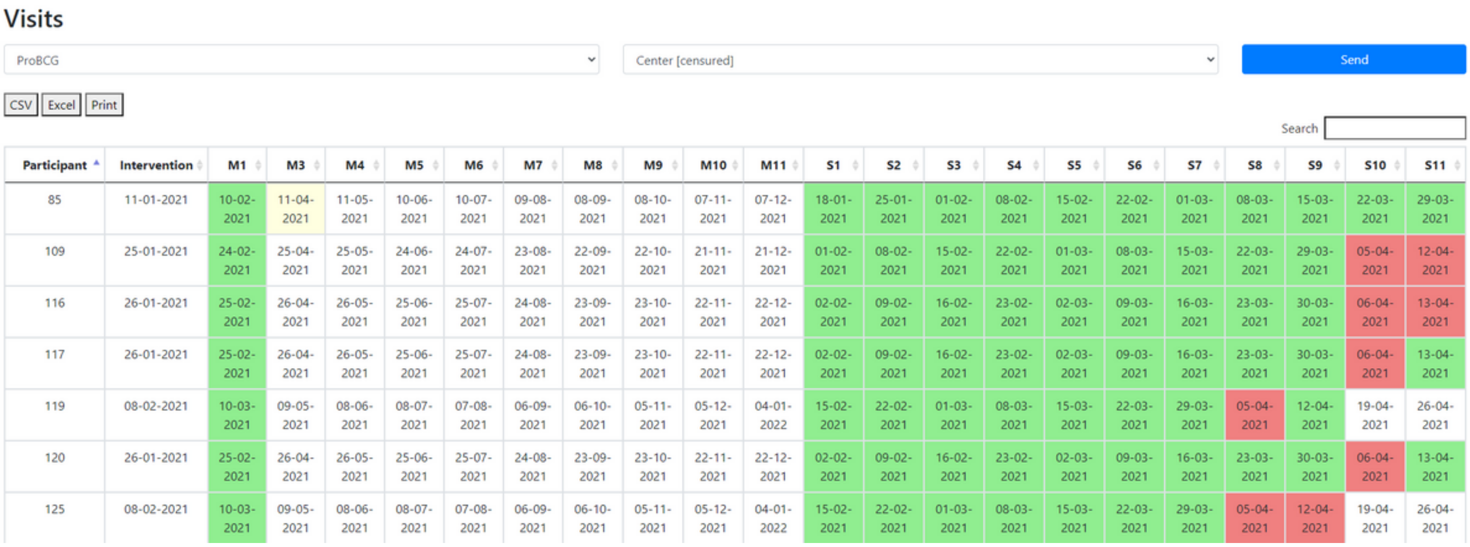


Validation Issues

RECORD ID	FORM	INSTANCE	FIELD	RULE	VALUE TO COMPARE	AUX FUNCTION	DATE	SOLVED?	DATE SOLVED	ACTION
								□		
79-192	Tratamento	1	tratamento_complete	Igual	0	complete_by_non_admin	Mar 28, 2021, 7:17:15 PM	!		✎
79-110	Identificação	1	identificao_complete	Igual	0	complete_by_non_admin	Mar 28, 2021, 7:34:40 PM	✓	Mar 30, 2021, 4:00:04 PM	✎
79-111	Identificação	1	identificao_complete	Igual	0	complete_by_non_admin	Mar 28, 2021, 7:34:42 PM	!		✎
79-153	Identificação	1	identificao_complete	Igual	0	complete_by_non_admin	Mar 28, 2021, 7:35:18 PM	✓	Mar 28, 2021, 8:21:16 PM	✎
79-190	Identificação	1	identificao_complete	Igual	0	complete_by_non_admin	Mar 28, 2021, 7:35:51 PM	!		✎
79-191	Identificação	1	identificao_complete	Igual	0	complete_by_non_admin	Mar 28, 2021, 7:35:53 PM	!		✎
79-192	Identificação	1	identificao_complete	Igual	0	complete_by_non_admin	Mar 28, 2021, 7:35:54 PM	!		✎
89-52	Identificação	1	data_1a_consulta_centro	Menor ou igual	2019-12-31		Mar 28, 2021, 7:38:56 PM	!		✎
89-52	Identificação	1	data_1a_consulta_espec	Menor ou igual	2019-12-31		Mar 28, 2021, 7:38:56 PM	!		✎
113-4	Identificação	1	identificador	CPF ou CNS (somente números)			Mar 28, 2021, 7:42:12 PM	!		✎

Figure 4

Data Quality Module – Validation issues dashboard



Visits

Participant	Intervention	M1	M3	M4	M5	M6	M7	M8	M9	M10	M11	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
85	11-01-2021	10-02-2021	11-04-2021	11-05-2021	10-06-2021	10-07-2021	09-08-2021	08-09-2021	08-10-2021	07-11-2021	07-12-2021	18-01-2021	25-01-2021	01-02-2021	08-02-2021	15-02-2021	22-02-2021	01-03-2021	08-03-2021	15-03-2021	22-03-2021	29-03-2021
109	25-01-2021	24-02-2021	25-04-2021	25-05-2021	24-06-2021	24-07-2021	23-08-2021	22-09-2021	22-10-2021	21-11-2021	21-12-2021	01-02-2021	08-02-2021	15-02-2021	22-02-2021	01-03-2021	08-03-2021	15-03-2021	22-03-2021	29-03-2021	05-04-2021	12-04-2021
116	26-01-2021	25-02-2021	26-04-2021	26-05-2021	25-06-2021	25-07-2021	24-08-2021	23-09-2021	23-10-2021	22-11-2021	22-12-2021	02-02-2021	09-02-2021	16-02-2021	23-02-2021	02-03-2021	09-03-2021	16-03-2021	23-03-2021	30-03-2021	06-04-2021	13-04-2021
117	26-01-2021	25-02-2021	26-04-2021	26-05-2021	25-06-2021	25-07-2021	24-08-2021	23-09-2021	23-10-2021	22-11-2021	22-12-2021	02-02-2021	09-02-2021	16-02-2021	23-02-2021	02-03-2021	09-03-2021	16-03-2021	23-03-2021	30-03-2021	06-04-2021	13-04-2021
119	08-02-2021	10-03-2021	09-05-2021	08-06-2021	08-07-2021	07-08-2021	06-09-2021	06-10-2021	05-11-2021	05-12-2021	04-01-2022	15-02-2021	22-02-2021	01-03-2021	08-03-2021	15-03-2021	22-03-2021	29-03-2021	05-04-2021	12-04-2021	19-04-2021	26-04-2021
120	26-01-2021	25-02-2021	26-04-2021	26-05-2021	25-06-2021	25-07-2021	24-08-2021	23-09-2021	23-10-2021	22-11-2021	22-12-2021	02-02-2021	09-02-2021	16-02-2021	23-02-2021	02-03-2021	09-03-2021	16-03-2021	23-03-2021	30-03-2021	06-04-2021	13-04-2021
125	08-02-2021	10-03-2021	09-05-2021	08-06-2021	08-07-2021	07-08-2021	06-09-2021	06-10-2021	05-11-2021	05-12-2021	04-01-2022	15-02-2021	22-02-2021	01-03-2021	08-03-2021	15-03-2021	22-03-2021	29-03-2021	05-04-2021	12-04-2021	19-04-2021	26-04-2021

Figure 5

Data Quality Module – Visits panel