Case Report Form based on semantic Web technologies

Ángel Esteban-Gil[1] and J. T. Fernández-Breis[2]*

[1]Fundación para la Formación e Investigación Sanitarias de la Región de Murcia, IMIB-Arrixaca-UMU, 30003 Murcia, Spain
angel.esteban@ffis.es
[2]Dpto. Informática y Sistemas, Facultad de Informática, Universidad de Murcia, IMIB-Arrixaca-UMU, 30100 Murcia, Spain
jfernand@um.es

Abstract.OBJECTIVE: Improving the capture, sharing and reuse of clinical research data within a biomedical research institute through the use of semantic case report forms (CRF).

BACKGROUND:Biomedical researchers need software solutions that allow working in projects with different, heterogeneous and changing information. A CRF is a set of questionnaires used for capturing the data of the patients recruited in a biomedical research study. Current CRF technological solutions have little flexibility to modify their structure to adapt to new requirements without major software changes, and they lack a well-defined model for the exploitation, generation of alerts or data quality assurance.

METHODS: Our approach divides the CRF building in two phases: (1) the definition of the data structure and the workflow to register these data, and (2) the recruitment process where the CRF captures the clinical information of each patient and the exploitation of the results of the biomedical project. OWL ontologies are employed for the formal CRF representation including the workflow of the patients recruited in the biomedical project. RDF repositories were used to store the questionnaire of each patient in every stage and SPARQL was used to exploit the semantic information.

RESULTS: In this work we present a web platform that incorporates the benefits of Semantic Web technologies to build, execute and exploit CRFs in biomedical projects. Our platform contains data of more than 14.000 patients recruited in more than 100 biomedical research projects running in our research institute.

CONCLUSION: Semantic Web technologies facilitate the construction of CRF platforms that meet the needs of biomedical researchers. We plan to improve the interoperability of the CRF data retrieval process by providing extracts compatible with standards such as HL7, CEN/ISO 13606 or OpenEHR.

Keywords: Biomedical Informatics, Semantic Web, Case Report Form, Ontology

adfa, p. 1, 2011.

1 Introduction

Biomedical researchers need software solutions able to exploit heterogeneous dynamic, project-specific information. A case report form (CRF) is a set of questionnaires used for capturing the data of each patient recruited in a biomedical research project [1]. However, heterogeneity is common in CRFs, because each study defines its own report schemas. More than 48000 clinical trials have been registered in Europe since 2014 [2], which means that such studies manage a large volume of information.

The Semantic Web can be seen as an extension of the current web, in which information is given well-defined meaning, better enabling computers and people to work in cooperation [3]. Ontologies [4] constitute the standard knowledge representation mechanism for the Semantic Web, and technologies such as OWL [5], RDF [6], and SPARQL [7] enable a formal representation of the domain, the data and their exploitation.

Many technological solutions are available to manage data for CRF nowadays [1], which can be grouped in two classes according to the type of technology used for representing and persisting the data: relational databases; and non-relational databases. The main disadvantage of the first approach is the little flexibility of the relational model for structural modifications without major changes in the software. The main disadvantage of the second approach is the lack of a well-defined model for exploitation, generation of alerts, or quality assurance of the data.

Our main objective is to develop a Web platform that facilitates the process of building and managing a CRF using Semantic Web technologies including: (1) the use of Semantic Case Reports Form for capturing the clinical data, and (2) the definition of customizable search interfaces and dashboards for the analysis and visualization of patient data.

In this context, our approach uses semantic web technologies for storing biomedical data in a flexible data model and exploiting it thanks to the semantic model that describes the data. Furthermore, these technologies permit the reuse of biomedical ontologies and the semantic interoperability of health resources when are required. Our approach has been applied so far in 10 biomedical projects and in 91 clinical trials, whose samples are stored the biobank in the Institute for Bio-health Research of Murcia (IMIB-Arrixaca-UMU) in Spain.

2 Methods

Our approach has two main stages (see Figure 1). The first stage is the definition of the data structures and the workflow that will be used for data capture. Workflows permit to determine the data capture stages included in the clinical study. The second stage is the execution of the CRF, which consists of capturing and exploiting the patient data.

One special feature of our approach is that the data manager can change the data structures and data workflow during the CRF execution. This feature provides flexibility to biomedical researchers, who can adapt their datasets to new requirements. This is enabled by the use of an OWL ontology[1], which can be extended with the specific concepts of the projects. This ontology has been built based on the existing ontologies like OBI[8], SIO[9] and SOPHARM[10]. The basic concepts of this ontology are described next:

- *Stakeholder*. The recruitment process of a biomedical study has several stakeholders: (1) the *Datamanager* represents the responsible of the patient data capture; (2) the *Researcher* stands for users who can capture and exploit the information, (3) the *Monitor* represents users who can monitor the captured patient data and the adverse effects in the study; and (4) the *Manager* represents users who promote the clinical study, can define the CRF and exploit its results. The manager is often the promoter or funder of the study, such as a pharmaceutical laboratory or a hospital.
- *Project.* The ontology includes a hierarchy of types of biomedical projects such as Clinical Trials, Observational Studies, Cohort Studies, etc.
- *Patient*. For each patient recruited, and due to the Spanish data protection law, we only capture gender, birth date and a code that ensures the anonymity of the patient. Each individual of the class *Patient* has a *Protocol* in the project. This permits, for instance, separating the sick individuals from the healthy ones.
- *Report.* This concept represents the set of information that must be captured in a concrete clinical interaction: the applied therapy, the results of a medical test, etc.
- *Stage*. This concept represents the stage or phase of a patient. For each stage the data manager can capture one or more *Reports*.

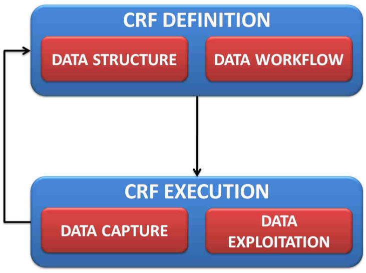


Fig. 1. Methodology schema

A Semantic Case Report Form is defined as an instance of Report that contains the answers to the items of the questionnaire, which are associated with a given Patient, which is in a concrete stage of its disease.

2.1 CRF Definition

The CRF definition has two phases: (1) the definition of the reports and (2) the definition of the workflow for each patient included in the biomedical project.

[1] http://www.imib.es/ontologies/CRDv4.owl

The generation of reports consists of defining the data capture fields. Our approach allows the definition of different types of fields: numbers, dates, times, text, boolean and enumerated. Enumerated fields permit to select and reuse, as values, classes from existing ontologies, including those used in other reports. All the fields and reports can be reused in different stages of the protocol or in different studies allowing the standardization of the information, so enabling its sharing and comparability.

When the data manager associates fields in a report, she can apply the following types of rules, which are implemented in the semantic model:

- **Cardinality rules.** They indicate the minimum and maximum cardinality for a given datum in this report. The cardinality can be a fixed value or it can be relative to the values of other fields. For example, a field "number of children" may affect the number of times age values for the children will have to be stored. Other example could be a field with the question "Do you smoke?". In the case of negative answer the field "number of daily cigarettes" could be null.
- **Range rules.** They indicate the range of values for a field in a report.
- **Format rules.** They are regular expressions to satisfy by the user when providing a value for this field. These rules are useful to store values such as emails, phone numbers, etc.

Our approach also permits to define derived fields. For example, if we have a field "weight" and other "height", the data manager can create a new field named "body mass index" calculated from the values of the previous fields.

The definition of workflows is based on state machines. In [11], the authors present a set of initial and end states and a set of intermediate states where the information transit from the initial state to the final one. When the data manager defines the states of the study, each one has a state machine associated. The configuration of each state requires the next information:

- **Transitions between states.** The transitions represent the path for the patient data in the recruitment process of a biomedical project.
- **People responsible of the data capture**.
- **Reports.** Which reports have to be filled in each state. For example, the screening phase of a clinical trial may include a report related to blood test, which will be used for recommending the next state for the patient.
- **Alerts.**
  - **Time alerts.** The patient is in the same state longer than expected.
  - **Data quality alerts.** The captured data are not enough to characterize the clinical features of the patient.

Figure 2 shows an example of a state machine. We can observe an initial state (blue), several end states (red) and a set of intermediate states related by transitions between them. Each state may have one or more responsible people, and have different alerts and reports associated.

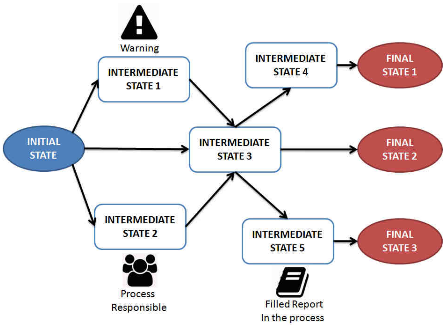


Fig. 2. An example of a general state machine used in our approach

Semantic transformation

The ontology that provides the basic knowledge entities is extended with entities of interest for the CRF. We initially proposed to the biomedical researchers the use of Protégé[2] for such purpose, which was rejected by them. Hence, we developed a web editor with features closer to their requirements and more intuitive for the intended users of the system. This also required developing a process to transform the content generated with this editor to OWL, what was done applying the following steps:

- Generation of the biomedical project and the different stakeholders that will complete and review the recruitment process in the CRF.
- Generation of the several protocols for each patient in each biomedical project.
- Creation of classes for each defined report. Each report has the fields as *owl:DatatypeProperty* and relationships as *owl:ObjectProperty* defined by the data manager. Each field maintains the rules of integrity, cardinality and range.
- Generation of the workflow of the CRF using our model of state machine. Besides the workflow generation, we need a class that represents the patient stage in each phase of the study. Another important aspect is the generation of the rules to transit between states, and the preconfigured alerts.

The ontology generated is stored in a document management system (DMS) with version control. The DMS helps us to exploit information captured in older versions of the recruitment protocol of the study. Our approach allows the user to choose the ontology version to exploit the clinical information stored in RDF.

2.2 CRF Execution

The CRF definition produces an OWL ontology that represents the structure of the data to be captured and the workflow to be applied to each patient recruited in the clinical study. Starting the recruitment requires to use our semantic running engine to

[2] http://protege.stanford.edu/

capture the data and our semantic exploitation model to take advantage of the information registered in the CRF.

Semantic running engine

The semantic running engine generates web forms for adding and updating the information of each semantic report, applying the rules defined in the report fields and in the state machines. The information is stored in a semantic repository with two types of data sources: (1) an OWL files server with the formal representation of the domains, and (2) an RDF repository which stores the data. We use Virtuoso[3] as data store. Virtuoso has been used in other effort such as [12].The ontologies guide all the layers of the solution: data capture, information delivery and exploitation.

The ODS (Ontology Driven-Searcher)[13] is the service for information delivery. This tool is an editor of SPARQL queries supported by OWL models. The tool uses the underlying CRF ontologies to show the necessary information to visually define SPARQL queries.

Semantic exploitation model

Our proposal includes a set of methods for exploiting the information stored in the semantic repository:

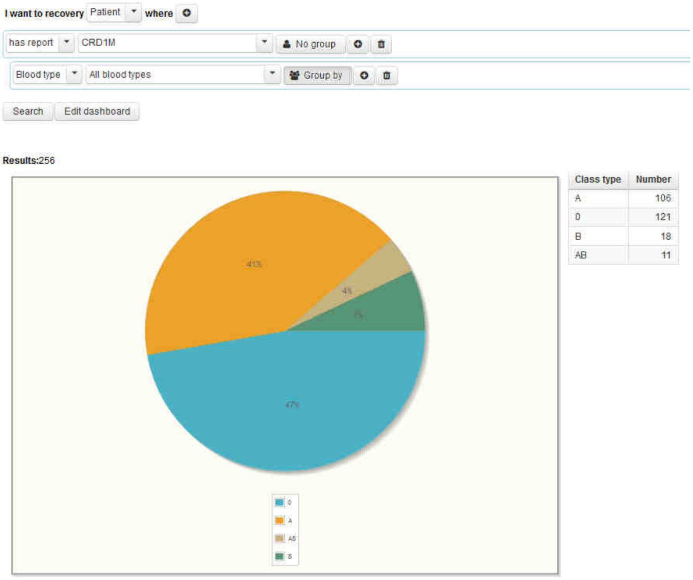
- **Semantic searcher**. This method uses the ODS for defining queries over the semantic CRF data model.
- **Alert management.** This method allows the generation of alerts over the semantic data. It uses the ODS for defining the alerts as queries and comparing the results with thresholds when these have been defined. For example, if the value of the systolic blood pressure is greater than 15, then the user may receive a high critical alert that implies that this patient is not suitable for the clinical study.
- **Semantic dashboard.** This tool permits users to formulate incremental, user-defined queries with a graphical user interface based on the ODS. The query results can be displayed in several customizable ways, allowing for the generation of on-demand dashboards.

3 Results

The approach described in the previous section has been applied in more than 100 biomedical projects in the Institute for Bio-health Research of Murcia (IMIB-Arrixaca-UMU). The platform is completely functional since January 2015. Our platform has had an important impact in our biomedical research institute. Nowadays we have the next results:

[3] `http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/`

- More than 14.000 patients have been recruited in several studies.
- More than 9.500 reports have been registered in the platform.
- More than 70 reports have been defined with more than 1.500 fields.
- More than 300 stages have been defined for the biomedical projects.
- The researchers have reused only 41 fields between reports.
- The researchers have reused only 10 reports between clinical studies.
- The platform has more than 50 users.
- Two CRFs involve patients from several regions of Spain.
- The researchers have configured 6 dashboards to exploit the data in real time. Figure 3 illustrates how the researcher uses the ODS to define a query over the semantic repository to represent graphically the blood type of the patients recruited in a biomedical project. This variable is captured in the report called "CRD1M".
- One biomedical project is using a standardized ontology, the ICD10 one[11].



I want to recovery Patient where

has report CRD1M No group

Blood type All blood types Group by

Results:256

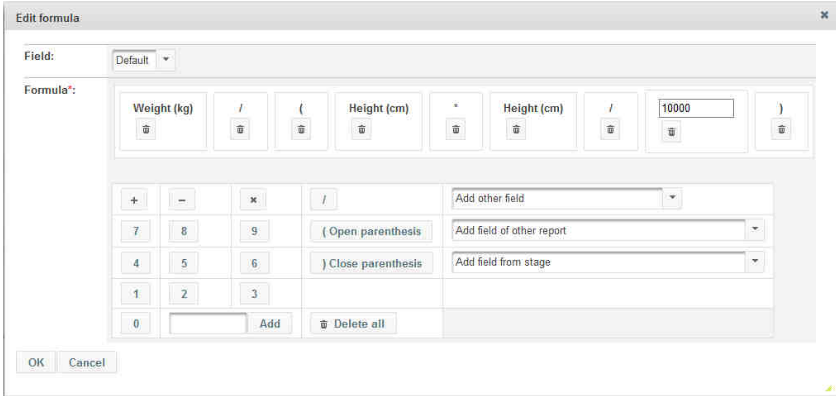
Class type	Number
A	106
0	121
B	18
AB	11

Fig. 3. Semantic dashboard

Thanks to the feedback of the researchers of the biomedical groups, we have included additional services to the platform:

- Generation of the patient's visit calendar from the information of the workflow of the clinical study. This calendar allows the generation of alerts for the clinical when they have to contact the patient.
- Use of a web calculator to define derived fields. An example that calculates the body mass index is shown in Figure 4.
- Use of the reports to characterize, not only patients, but also their biological samples, such as data from the pathological analysis of a tumor.
- Filling of the reports from mobile devices such as smartphones or tables.

- Generation of PDF reports with the results of the patients saved in the CRF.



Edit formula

Field: Default

Formula*: Weight (kg) / (Height (cm) * Height (cm) / 10000)

Fig. 4. Definition of a derived field

4 Discussion and conclusions

In this work we have demonstrated that the semantic technologies are able to manage and exploit the results of the patients' recruitment process in any clinical study. We have evaluated this tool following the method proposed in [1], obtaining with good results in comparison to other tools available in the market. Our platform fulfills all criteria except the capability to export the data in CDISC format [15].

Many researchers have integrated semantic web technologies in biomedical research. We have grouped these proposals in two types: (1) the use of ontologies to classify the clinical information [16], and (2) enrichment of biomedical data for exploiting using semantic technologies [17]. Our approach is different because we resolve the problem from a global perspective, trying to use semantic technologies in the whole life cycle of the biomedical project. The main advantage of our proposal is that ontologies guide all the process of the biomedical project including the capture of data and its exploitation. Furthermore, the logical schema of the CRF may help to understand the recruitment process and the results of the study.

Our approach shows how semantic web technologies permit researchers to adapt the CRF to their specific requirements without the help of and IT expert. The use of an RDF repository allows for building a robust and scalable architecture for big clinical data warehouses [12]. Furthermore, this architecture is very flexible in changing environments as the biomedical research. The use of OWL ontologies to represent the knowledge stored in the RDF repository allows to exploit it using technologies such as the ODS to define queries without mastering semantic technologies. Another important benefit in the use of OWL is the capability to reuse the fields and concepts among several projects and to take advantage of the clinical knowledge modeled in this format.

We have learned from the use of the platform that some users are only interested in the exploitation of data, so we are developing data retrieval methods for importing data from other systems they are currently using to capture patient data.

Finally, this platform has also an economic impact in our organization. In the last 6 years, IMIB-Arrixaca-UMU has run 12 clinical trials funded by industry. Our researchers have used paper-based CRF in six of them. The cost of the electronic CRF used in the other six clinical trials was almost 55.000 € (9.000 € on average). We are not able to calculate the effort to exploit the paper-based CRF data, but our electronic CRF platform, which has been used in ten non-industrial clinical trials, has permitted to save approximately 90.000 €.

Our approach presents some limitations: (1) our solution is not able to automatic retrieve data from other clinical systems, (2) our solution does not implement any clinical standard to interoperate with other clinical software, (3) we have not been able to convince researchers to publish and share their questionnaires,(4) the researchers still prefer CSV data exploitation instead of using our semantic exploitation model, (5) our reuse of biomedical ontologies is still limited and (6) we are exploiting OWL reasoning yet.

As future work we plan to improve the interoperability between our CRF and other clinical systems implementing standards as HL7, CEN/ISO 13606 or openEHR. We also plan to export the data to CDISC [15]. We are planning to incorporate ontology alignment techniques to improve the reuse and standardization of our CRF semantic models. Finally, we plan to provide training to promote the use of the semantic exploitation model.

To conclude, the construction of tools that facilitate the use of standard clinical terminologies or the reuse of fields or reports will improve the exploitation of the data aggregated of several patients included in different studies achieving the desired real goal in the biomedical research: improving health care to the patient.

Acknowledgments

This project has been possible thanks to the cooperation of the clinical research groups of the Institute for Bio-health Research of Murcia (IMIB-Arrixaca-UMU). This work was supported by the Ministerio de Economía y Competitividad and the FEDER programme through grant TIN2014-53749-C2-2-R2, and the Fundación Séneca through grant 19371/PI/14.

References

1. Leroux, H., McBride, S., Gibson, S. On selecting a clinical trial management system for large scale, multi-centre, multi-modal clinical research study., in: HIC. pp. 89–95 (2011).
2. European Medicines Agency. European Clinical Trials Database (EudraCT). `https://eudract.ema.europa.eu/`. (accessed September 2016)
3. Berners-Lee, T., Hendler, J., &Lassila, O.: The semantic web. Scientific American, 284, 34–43(2001).

4. Studer, R., Benjamins, V. R., &Fensel, D.: Knowledge engineering: Principles and methods. Data & Knowledge Engineering, 25, 161–197(1998).
5. W3C, OWL2 web ontology language. `http://www.w3.org/TR/owl2-overview/` (accessed September 2016)
6. W3C, RDF Resource Description Framework. `http://www.w3.org/RDF/` (accessed: September 2016)
7. W3C, SPARQL Query Language for RDF. `http://www.w3.org/TR/rdf-sparql-query/` (accessed: September 2016)
8. Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, Clancy K, Courtot M, Derom D, Dumontier M, Fan L, Fostel J, Fragoso G, Gibson F,Gonzalez-Beltran A, Haendel MA, He Y, Heiskanen M, Hernandez-Boussard T, Jensen M, Lin Y, Lister AL, Lord P, Malone J, Manduchi E, McGee M, Morrison N, Overton JA, Parkinson H, Peters B, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Schober D, Smith B, Soldatova LN, Stoeckert CJ Jr, Taylor CF, Torniai C, Turner JA, Vita R, Whetzel PL, Zheng J. The Ontology for Biomedical Investigations. PLoS One. 2016 Apr 29;11(4):e0154556.
9. Dumontier M, Baker CJ, Baran J, Callahan A, Chepelev L, Cruz-Toledo J, Del Rio NR, Duck G, Furlong LI, Keath N, Klassen D, McCusker JP, Queralt-Rosinach N, Samwald M, Villanueva-Rosales N, Wilkinson MD, Hoehndorf R. The Semantic science Integrated Ontology (SIO) for biomedical research and knowledge discovery. J Biomed Semantics. 2014 Mar 6;5(1):14.
10. Coulet A, Smaïl-Tabbone M, Napoli A, Devignes M. Suggested Ontology for Pharmacogenomics (SO-Pharm): Modular Construction and Preliminary Testing. Proceedings of International Workshop on Knowledge Systems in Bioinformatics – KSinBIT'06, Oct 2006, Montpellier, France. 2006.
11. Latella D, Majzik I, Massink M (1999) Towards a formal operational semantics of UML statechart diagrams. In: Formal Methods for Open Object-Based Distributed Systems, Springer. pp. 331_347.
12. Rea, S., Pathak, J., Savova, G., Oniki, T., Westberg, L., Beebe, C., Tao, C., Parker, C., Haug, P., Huff, S., Chute, C.: Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project, Journal of Biomedical Informatics. 45, 736-771 (2012).
13. Esteban-Gil, A., Garcia-Sanchez, F., Valencia-Garcia, R., Fernandez-Breis, JT.: Social-BROKER: A collaborative social space for gathering semantically-enhanced financial information, Expert Systems with Applications. 39, 9715-9722 (2012).
14. Cardillo, E., Tamilin, A., Eccher, C., Serafini, L.: ICD-10 Ontology, `https://dkm.fbk.eu/technologies/icd-10-ontology` (accessed September 2016)
15. CDISC. Clinical Data Interchange Standards Consortium. Available from: http://www.cdisc.org/standards.(accessed September 2016)
16. Leroux, H., Lefort, L. Semantic enrichment of longitudinal clinical study data using the CDISC standards and the semantic statistics vocabularies. Journal of biomedical semantics 6, 1 (2015).
17. Richesson, R.L., Andrews, J.E., Krischer, J.P., 2006. Use of SNOMED CT to represent clinical research data: a semantic characterization of data items on case report forms in vasculitis research. Journal of the American Medical Informatics Association 13, 536–546.