Journal of Biomedical Informatics 63 (2016) 157–168

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Clinical Trial Information Mediator ☆

Christian Krauth [a,*], Wolfgang Kuchinke [a], Martin Eckert [a], Rene Bergmann [a], Benjamin Braasch [a], Töresin Karakoyun [c], Christian Ohmann [b]

[a] Clinical Research Informatics, Coordination Centre for Clinical Trials, Heinrich-Heine-University, Moorenstr. 5, 40225 Duesseldorf, Germany
[b] ECRIN, Kaiserswerther Str. 70, 40477 Duesseldorf, Germany
[c] Bayer Pharma AG, Aprather Weg 18, 42096 Wuppertal, Germany

ARTICLE INFO

Article history:
Received 13 April 2016
Revised 4 July 2016
Accepted 7 August 2016
Available online 8 August 2016

Keyword:
Clinical Trial Information Mediator (CTIM)
Database
Integration
BioMedBridges
PubMed
BioSample
ClinicalTrials.gov

ABSTRACT

For research in biomedical sciences, cross-domain searches through several different databases are an increasingly necessary task that often becomes a time consuming and labour-intense process. This is especially the case when different domain databases have to be combined, for example combined searches in clinical trials registries, publication databases and research databases. The Clinical Trial Information Mediator (CTIM) addresses this problem and offers a novel way for the combined search in ClinicalTrials.gov, PubMed and BioSamples.

CTIM was developed based on a requirements analysis and implemented using open source technology. A search engine with a graphical user interface was developed in order to search linked data in the three databases ClinicalTrials.gov, PubMed and BioSamples; thereby enabling CTIM to bridge the gap between different knowledge domains of clinical trials, publications of research results and biosamples/genetic information. CTIM was applied in three use cases demonstrating that information retrieval could be considerably improved in sense for complex queries. These use cases show that more relevant results were obtained and more associated publications and biosamples could be retrieved in comparison to a separate single search.

Main advantages of CTIM are identifying related information between clinical trials and publications employing a clinical trial centred kind of search, simplified access to its databases and thus reduced search time. In addition it can be used by researchers without prior training because of the intuitive usage.

1. Introduction

Research in life sciences progressively requires cross-domain searches through different databases. This is often a time consuming process due to querying different databases separately using different user interfaces [13]. Especially researchers interested in information about clinical trials with a view to design new studies must find trial related information in different biomedical databases. The first step in the query process hereby is to express a well-formulated research question [1] that may cover questions about the characteristics of the population, the intervention and the outcome of the study. The main databases used for literature searches for biomedical research are MEDLINE® [27] that contains journal citations and abstracts of biomedical literature and PubMed® [28,43] that provides free access to MEDLINE and links to many full text articles. A number of web tools that provide comparable literature search services for PubMed have been developed and were evaluated [23], such as Medline complete [9], BioCreative [3], RefMed [30], MEDPILOT [48], iPubMed [42], CAIPIRINI [38], to name but a few.

Clinical trials registers such as ClinicalTrials.gov [26] and the World Health Organization (WHO) International Clinical Trials Registry Platform (ICTRP) [44] are increasingly used to identify ongoing or already finished clinical trials [15]. These registers offer important information on the methods and progress of trials. ClinicalTrials.gov runs by the United States National Library of Medicine (NLM) at the National Institutes of Health and is the largest register of its kind. It contains registrations from over 200,000 trials in more than 170 countries. Data in ClinicalTrials.gov are self-reported by trial sponsors or investigators [19]. Each record

☆ **Availability of the tool:** The current pilot can be accessed via: http://hhu2.at.xencon.de:8080/ctim/. The tool is licensed under the Apache Licence 2.0 and the code can be provided upon request.

* Corresponding author.
E-mail addresses: christian.krauth@uni-duesseldorf.de (C. Krauth), kuchinke@med.uni-duesseldorf.de (W. Kuchinke), martin.eckert@uni-duesseldorf.de (M. Eckert).

http://dx.doi.org/10.1016/j.jbi.2016.08.012
1532-0464/

contains a set of mandatory data elements that describe the study's purpose, recruitment status, study design, eligibility criteria, and sites involved. A trial with an identification number that is registered in ClinicalTrials.gov can be linked to a journal article with a PubMed identification number [47]. ClinicalTrials.gov offers different search options. In the basic search option one or more terms are entered into a single search box. Advanced search means that search terms in multiple fields for items such as recruitment status and location are entered or selected.

In addition, data in many other research databases or repositories are important for research questions [25]. The European Bioinformatics Institute (EBI) [36] is a centre for research and services in bioinformatics and hosts a number of these publicly available open life science resources, including biomedical databases, analysis tools and ontologies. The most important ones are ArrayExpress (gene expression), ChEBI (Chemical Entities of Biological Interest), ENA (European Nucleotide Archive), Ensembl (genome databases for vertebrates and other eukaryotic species), Europe PubMed Central (biomedical research literature), MetaboLights (metabolomics data), and Protein Data Bank (European resource for the collection, organization and dissemination of data on protein structures).

The BioSamples Database (BioSD) [17] stores information about biological samples used in gene sequencing. Examples for biosamples include a primary tissue biopsy, an individual organism or an environmental isolate. The BioSamples database captures sample metadata in a structured way by encouraging use of controlled sample attribute field name vocabularies.

BioMedBridges (BMB) [11] is an EU-funded project that aimed to develop harmonised solutions to support data sharing between research infrastructures (RI) in the field of life sciences. These harmonised solutions consist of registries, standards, best practices and interoperability tools. Data sharing has to cover interoperability, semantic integration, and data security. BioMedBridges (BMB) approach allows interoperability between data and services in the biological, molecular, medical, translational and clinical fields, thereby connecting very heterogeneous data sources, like sequencing data, bio-samples data, clinical data, spectroscopy and imaging data and even synchrotron data for structure determination, in order to be able to address research questions on a broader data basis than has been done before.

The formulation of an effective research question is one of the most important steps in research [1]: it is necessary for the researcher to have some knowledge to be able to generate an idea and to formulate a question. Based on this prior knowledge and the data the researcher already has, the researcher generates a hypothesis and formulates a research question to obtain meaningful data. In this way, the formulation of the research questions can lead to the need for more data, the enrichment of existing data, and perhaps the necessity to use a new data bridge (Fig. 1). CTIM was developed to support the generation of research questions for the field of clinical research.

In one of BMB's working packages, the pilot integration of REST based vignette services was established and drug information was integrated with information from clinical trials [12]. From this pilot the basic requirements for a tool to support the search of related entries in different databases were derived. Such a tool should link information about publications, genes, and drugs with the ones of clinical trials with the aim to provide more insight into the effect of drugs and genes and to design better clinical trials and in this to support the research process.

In particular, an intuitive user interface for a simple query system that links searches in clinical trial registries with relevant publications and information about samples/genes and drugs was seen as useful. For this purpose, clinical trials information from ClinicalTrials.gov, publication information from PubMed and biosamples data from the biobank registry BioSamples [37] were combined to generate linked data [45] and information that is provided in a single step, information, which would be more difficult to obtain by searching and analysing each database separately.

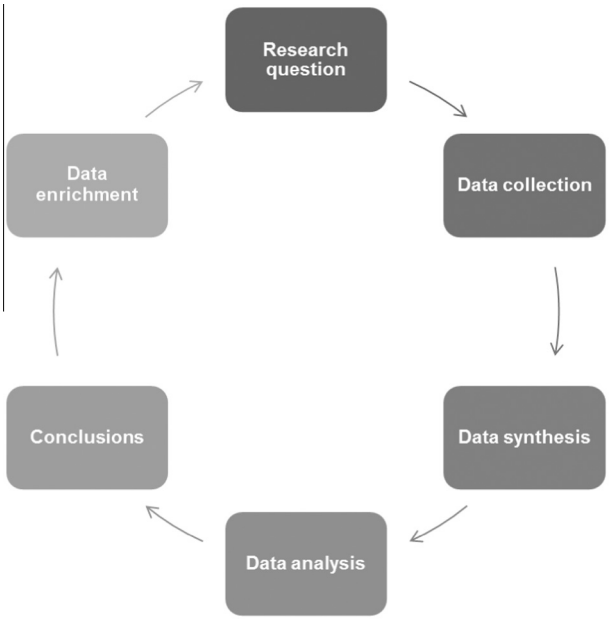

Fig. 1. General research process for the enrichment of data. The collection of data is the first step after a research question has been formulated. It is followed by the synthesis and analysis of the data. Through the analysis process the researcher can come to a conclusion that is mirrored in an enrichment of the data. This enriched data possibly leads to a new research question.

Finally, all the used database sources are available to all requesters, free of charge and encourage an approach to employ their data by providing guidance and suitable programmatic interfaces to their databases.

2. General approach

Based upon the workflow that generates research questions and is using database access to enrich data available about clinical trials, the approach of CTIM combines data collection, synthesis and analysis (Fig. 1) and in this way supports researchers to get additional data suitable to answer their research questions.

From the gathered requirements, three main requirements led the process of developing the tool:

- Focus on trial centred search.
- Simplified access with an intuitive user interface.
- Reduced query time, immediate generation of query results.

2.1. Trial centred search

CTIM implies that the researcher is starting a search by identifying a clinical trial in a clinical trial registry and then tries to enrich this information with further information from other data sources (e.g. publications or biosamples). For this purpose, the data of ClinicalTrials.gov was used as the initial data source for CTIM.

In general, study information is submitted to ClinicalTrials.gov at the beginning of a study, before the recruitment of participants' starts, and this information on the trial is updated throughout the study [47]. Each ClinicalTrials.gov record of a study displays summary information about a study protocol including, disease or condition, intervention, title and several further information.