FP7-ICT-611140 CARRE

Project co-funded by the European Commission
under the Information and Communication Technologies
(ICT) 7th Framework Programme

D.3.4. Aggregators for Medical Scientific & Educational Data

E. Liu, G. Gkotsis, H. Wei, X. Zheng, N. Portokallidis, G. Drosatos

January 2015

CARRE Contacts

Project Coordinator: Eleni Kaldoudi kaldoudi@med.duth.gr

DUTH Democritus University of Thrace	Eleni Kaldoudi	kaldoudi@med.duth.gr
OU The OpenUniversity	John Domingue	john.domingue@open.ac.uk
BED: BedfordshireUniversity	Enjie Liu	Enjie.Liu@beds.ac.uk
VULSK: Vilnius University Hospital SantariskiųKlinikos	Domantas Stundys	Domantas.Stundys@santa.lt
KTU Kaunas University of Technology	Arunas Lukosevicius	arunas.lukosevicius@ktu.lt
PIAP Industrial Research Institute for Automation & Measurements	Roman Szewczyk	rszewczyk@piap.pl

Disclaimer

CARRE is a Specific Targeted Research Project partially funded by the European Union, under FP7-ICT-2013-10, Theme 5.1. “Personalized health, active ageing & independent living”.

Document Control Page

Project	
Contract No.:	611140
Acronym:	CARRE
Title:	Personalized Patient Empowerment and Shared Decision Support for Cardiorenal Disease and Comorbidities
Type:	STREP
Start:	1 November 2013
End:	31 October 2016
Programme:	FP7-ICT-2013.5.1
Website:	http://www.carre-project.eu/

Deliverable	
Deliverable No.:	D.3.4
Deliverable Title:	Aggregators for medical scientific and educational data
Responsible Partner:	BED – Enjie Liu
Authors:	E. Liu, G. Gkotsis, H. Wei, X. Zheng, N. Portokallidis, G. Drosatos
Input from:	All partners
Peer Reviewers:	George Gkotsis (OU), Darius Jegelevicius (KTU)
Task:	T.3.4. Aggregators for medical scientific and educational data
Task duration:	10 months: 1 April 2014 to 31 January 2015
Work Package:	WP3: Data and metadata harvesting
Work Package Leader:	KTU – Arunas Lukosevicius
Due Date:	31 January 2015
Actual Delivery Date:	3 February 2015
Dissemination Level:	PU
Nature:	P
Files and format:	Deliverable report: 1 pdf file Software source code available from project web site, see Annexes 1-3: (1) CARRE_D.3.4_Aggregators_EvidenceData_Software_KnownRisk.7z (2) CARRE_D.3.4_Aggregators_EvidenceData_Software_NewRisk.zip.zip (3) http://carre.kmi.open.ac.uk (4) CARRE_D.3.4_Aggregators_MedicalEvidence_Educational_v0.2.zip
Version:	06
Status:	☐ Draft ☒ Consortium reviewed ☒ WP leader accepted ☒ Coordinator accepted ☐ EC accepted

Table of Contents

List of Figures

List of Tables

Document Revision History

Version	Date	Modifications	Contributors
v01.0	19 Jan 2015	template and content as in 1st year report	Eleni Kaldoud, Nick Portokallidis
v02.0	27 Jan 2015	Second version	Enjie Liu, George Gkotsis, Nick Portkalidis, Hui Wei, Xin Zheng
v03.0	29 Jan 2015	Third version	Enjie Liu, Nick Portokallidis, George Drosatos
v04	29 Jan 2015	reviewers' comments addressed, final editing for uniformity	Enjie Liu, E. Kaldoudi
v05	02 Feb 2015	editing and formatting, minor corrections	Hui Wei, Enjie Liu
v06	02 Feb 2015	formatting for QA conformance, Annexes	E. Kaldoudi

Executive Summary

This deliverable contains detailed information of aggregating medical evidence data and educational data as well as appropriate data aggregator architecture and implementations. The tasks are based on the previous analysis of risk associations of patients with cardiorenal diseases. The developments of the aggregators can be demonstrated through the use cases. The medical evidence data aggregator is designed to find additional evidence for the known risk associations as well as identify possible new risk associations which will be further evaluated by medical experts. Educational resource aggregator is to harvest educational resources from 3rd party repositories and present these to the medical expert for annotation and rating. The deliverable also describes the development of the Risk Model Semantic Data Entry system that is used for integrating and inputting of all medical evidence data and educational data into CARRE semantic repository.

About CARRE

CARRE is an EU FP7-ICT funded project with the goal to provide innovative means for the management of comorbidities (multiple co-occurring medical conditions), especially in the case of chronic cardiac and renal disease patients or persons with increased risk of such conditions.

Sources of medical and other knowledge will be semantically linked with sensor outputs to provide clinical information personalised to the individual patient, to be able to track the progression and interactions of comorbid conditions. Visual analytics will be employed so that patients and clinicians will be able to visualise, understand and interact with this linked knowledge and take advantage of personalised empowerment services supported by a dedicated decision support system.

The ultimate goal is to provide the means for patients with comorbidities to take an active role in care processes, including self-care and shared decision-making, and to support medical professionals in understanding and treating comorbidities via an integrative approach.

Terms and Definitions

The following are definitions of terms, abbreviations and acronyms used in this document.

Term	Definition
CSS	Cascading Style Sheets
DCMI	Dublin Core Metadata Initiative
EC	European Commission
eHealth	Electronic health
EU	European Union
ICD	International Classification of Diseases
ICT	Information and communication technologies
LOM	Learning Object Metadata
MedLinePlus	The National Institutes of Health's Web site for patients and their families and friends, http://medlineplus.gov.
PubMED	Free search engine accessing primarily the MEDLINE database www.ncbi.nlm.nih.gov/pubmed/
RDF	Resource Description Framework
RMSDE	Risk Model Semantic Data Entry
SPARQL	SPARQL Protocol and RDF Query Language
UMLS	Unified Medical Language System
Wikipedia	The free encyclopaedia that anyone can edit http://en.wikipedia.org/

1. Introduction

Task 3.4 involves the development of aggregators for medical evidence data and patient educational content from on-line authoritative sources. Aggregators will be built for all identified sources (T.2.3) of this type. These kinds of information are either openly available to public, such as some government medical advice sites, or access based on subscriptions, such as PubMED[1], and MedLinePlus[2].

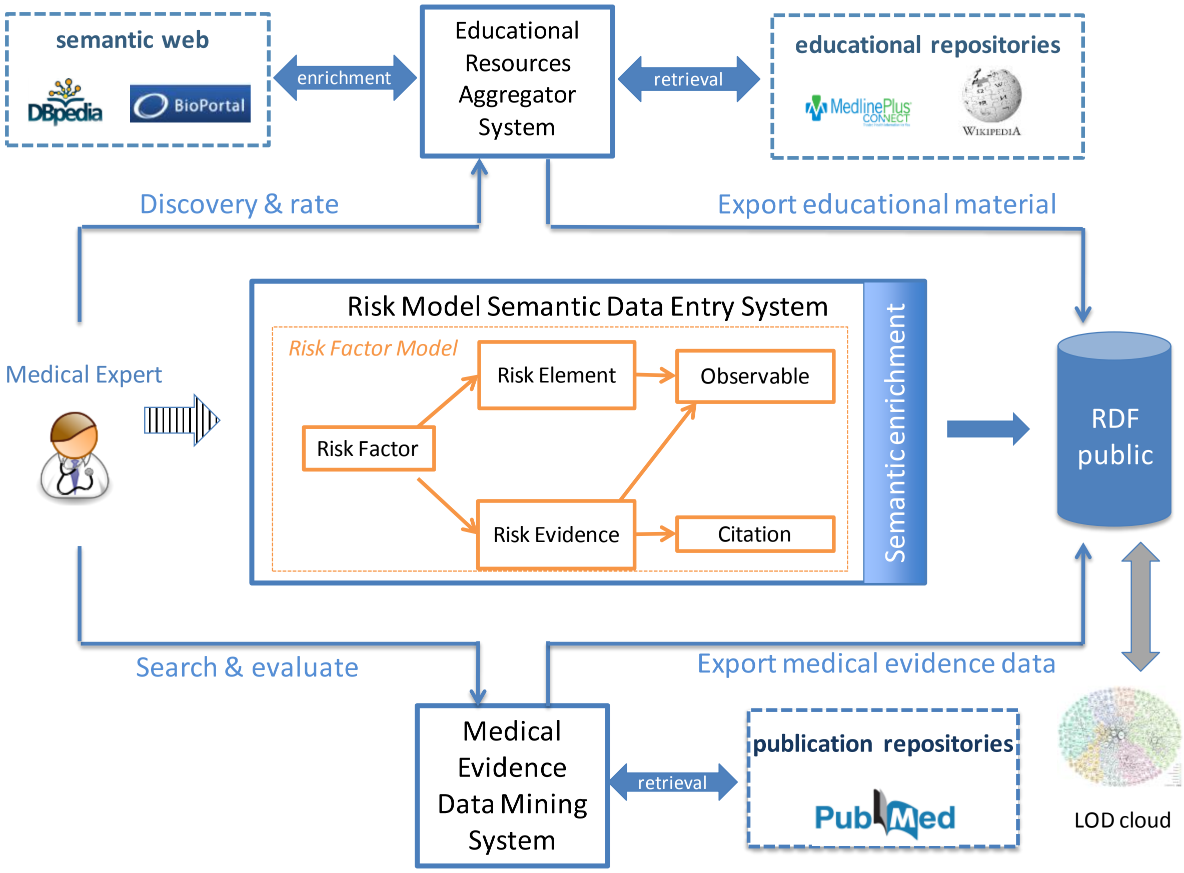

Figure 1. Functional overview of medical evidence aggregators described in D.3.4.

The aim of medical evidence data aggregators are to:

1) harvest data from medical state-of-the-art scientific literature databases for additional evidence on risk associations of cardiorenal disease and its comorbidity;

2) find new risk associations from latest publications of clinical trial results.

The task aims at achieving functions required for use case 10 & 11 defined in D2.1 Domain Analysis & Use Cases. The medical knowledge concerning the risk factors are obtained from D2.2 Functional Requirements & CARRE Information Model. The aim of the educational resource aggregator is to harvest educational resources from 3rd party repositories, present these to the medical expert for annotation and rating. Both aggregators will output the results of the annotation (together with resource metadata) to the CARRE public RDF repository.

---

1 PubMed, US National Library of Medicine, National Institutes of Health , http://www.ncbi.nlm.nih.gov/pubmed

2 MedLinePlus, Trusted Health Information for the Public, US National Library of Medicine, http://www.nlm.nih.gov/medlineplus/

1.1. Related use cases

Tasks in T3.4 aim to implement use cases 10 & 11 defined in D.2.1: Domain Analysis & Use Cases (see table 1 and 2).

Table 1. Use case 10	
ID	UC_PE_10
Title	Educational material based on current state and risks
Goal	The goal of this use case is to inform patients about their current health status and their risks.
Domain	Personalized Education
Description	In this use case, users have to insert medical data to CARRE system and then the system will analyse their data and send to them a feedback text with educational material based on the individual health state.
Participants	P1, P2, P3, P4, D1, D2
Pre-conditions	End users must input personal information.
Post-conditions	N/A

Table 2. Use case 11	
ID	UC_PE_11
Title	New medical evidence available
Goal	The goal of this use case is to inform end users about new medical evidence available about their wises
Domain	Personalized Education
Description	In this use case users enter the health condition that they are interested in and the system will support end users with a feedback text with the latest available medical evidence.
Participants	P1, P2, P3, P4, D1, D2
Pre-conditions	N/A
Post-conditions	N/A

1.2. Risk associations in CARRE

The risk associations are identified in D2.2: The medical Functional Requirements & CARRE Information Model. Here we give a brief summary.

The basic concepts in modelling comorbidity are:

- risk factor;
- risk association;
- risk element;
- observable; and
- evidence source.

Risk Element: Risk factor is the (often causal) association of an agent (*source risk element*) to a negative health outcome (*target risk element*). In cardiorenal disease and comorbidities, most often the (causal) agent is in itself a negative health outcome. In this sense, risk agents and their outcomes can be seen as instances of the same entity, called here '*risk element'*. Risk elements include all the disorders/diseases involved in the comorbidity under discussion as well as any other risk causing agent.

Risk Association: The association of one risk element as the risk source with another risk element, which is the negative outcome under certain conditions, is a '*risk association'*. This association is a rather complex one and is characterised by a number of other concepts:

Table 3 shows a sample risk factor and table 4 shows the support evidence. A full list of risk associations is in D2.2.

Table 3. Sample definitions of the risk factors	
Risk Factor	
Risk Source:	Acute kidney injury
Risk Target:	Chronic kidney disease
Association Type:	Is issue in
RiskID:	REID1
Author	Laurynas

Table 4. Sample definitions of the risk evidence	
Risk evidence ID1	
RiskID:	REID1
(Bio)marker:	Serum creatinine
Biomarker Condition:	-
Ratio Type:	HR
Ratio Value:	8.8
Confidence Interval:	3.1-25.5 (95%)
Adjusted for:	-
Evidence source:	Coca SG, Singanamala S, Parikh CR. Chronic kidney disease after acute kidney injury: a systematic review and meta-analysis. Kidney Int. 2012 Mar;81(5):442-8.
Evidence source PMID	PMC3788581
Evidence source type:	Systematic review and meta-analysis
Author	Laurynas

The rest of the report is arranged as follows. Section 2 gives general background of the aggregation of medical evidence data. Section 3 explains the development of aggregating further evidence for the known risk associations. Section 4 describes the approaches that are used in mining the unknown risk associations. Finally, Section 5 details the CARRE manual knowledge input system; section 6 dedicates to aggregation of educational data. Annexes 1-3 give links for downloading the actual D.3.4 deliverable which is the software for medical literature and educational resource aggregators developed in T.3.4 (and described in detail in this report).

2. Aggregators for medical evidence data

The purpose of the task is to gather medical knowledge with the aims to 1) enrich the evidence of the existing risk descriptions as entered manually by medical experts and 2) identify new risk associations for cardiorenal diseases and comorbidity as published in medical literature during and beyond the project's lifetime. This aggregator extracts and summarises key information from popular and trusted medical publications as they are indexed in PubMED.

Figure 2 shows the overall architecture of the aggregator. Some of the functions will be implemented as part of other tasks later in the project. At the frontend, a user interface is provided for the medical expert to test and validate the retrieved medical evidence data. It provides links to tools for data mining, data visualisation and evaluation.

At the backend, there are three main functions:

- key sentences extraction is the core component which uses data mining approaches to automatically identify medical evidence data;
- data process provides support functions for interactive risk association analysis which enables users to explore the identified evidence; it also provide functions for mapping the newly identify data onto CARRE schema and interactive with CARRE semantic data entry system;
- resource rating module will be used for group of medical professionals to evaluate and finalise the newly founded evidence.

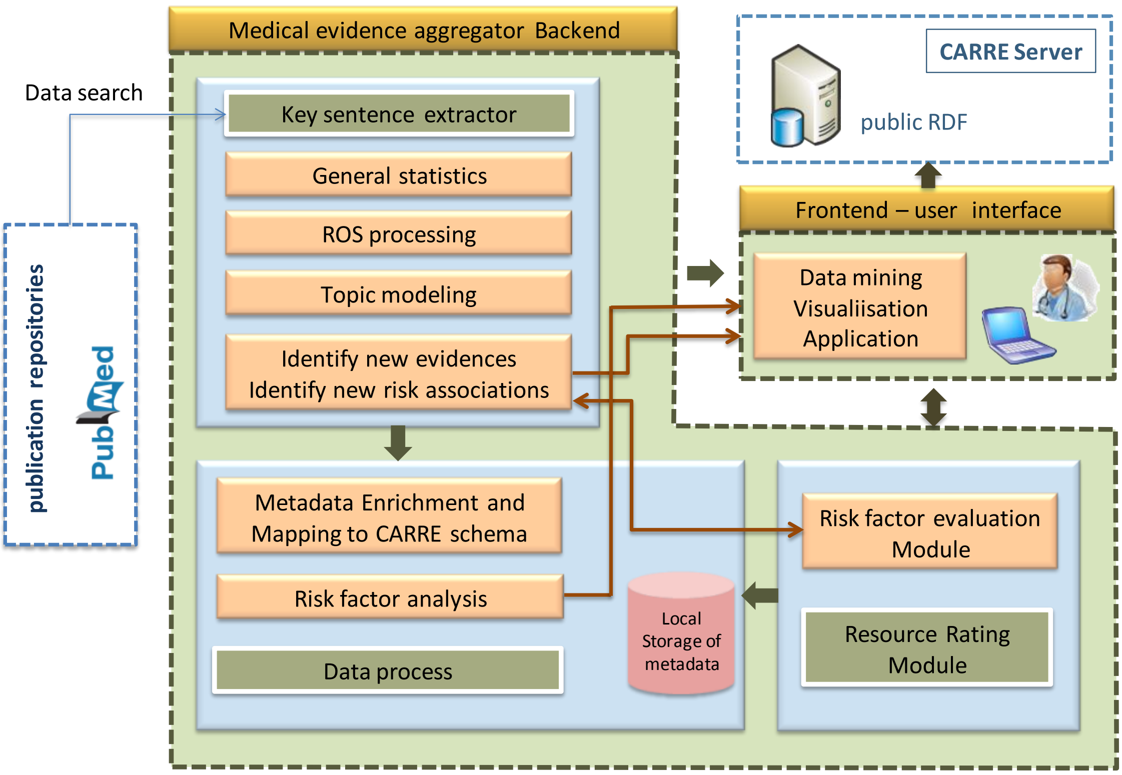

Figure 2. Overall architecture of aggregating medical evidence data.

2.1. Task breakdown

To support the use cases listed in section 1.1 the following four steps, shown in Figure 3, are identified. The description of the steps also maps them onto the functional blocks shown in Figure 2.

Step 1 data source search: The search starts from using PubMED APIs to retrieve the title and abstract (due to access control we cannot access full paper in some cases).

Step 2 data mining for medical evidence data: This subtask uses functions in functional blocks of 'key sentence extractor', 'data process' and 'resource rating module', as shown in Figure 2. This is the core step.

Step 3 analysis and evaluate risk factors and evidence: This subtask uses functions in the functional blocks of 'resource rating module', and 'risk factor analysis' of the 'data process'. This subtask is built for users to check and confirm the risk associations on a semi-automatic basis.

Step 4 output to semantic repository: This subtask is to output the identified risk factors to the CARRE RDF data repository. The Risk Model Semantic Data Entry system (RMSDE) – is an interface used by medical experts to record risk associations and its evidences identified in D.2.2.

Both step 1 and 3 will be carried out in a frontend user interface, as shown in Figure 2, which is a CARRE data-mining portal. Individual users can evaluate the identified evidence for step 3 using the portal, the group evaluation and voting for the evidence will be developed along with Task 5.1: Interactive visual interface, where visualisation tools will be provided for risk associations which can help group members in better understanding the current evidences and hence to make more accurate evaluations.

Step 2 focuses on data mining and in 2.2 the technical background is provided. Section 3 and 4 are used to describe the approaches used in T3.4.

Currently, step 4 is a manual input system for evidence data through data mining, and it can be undertaken automatically or semi-automatically together with the Task 4.3 in DOW: schema mapping & metadata enrichment. The semantic data entry system RMSDE is detailed in section 5.

The CARRE server will integrate the CARRE services, and this will take place at the integration stage.

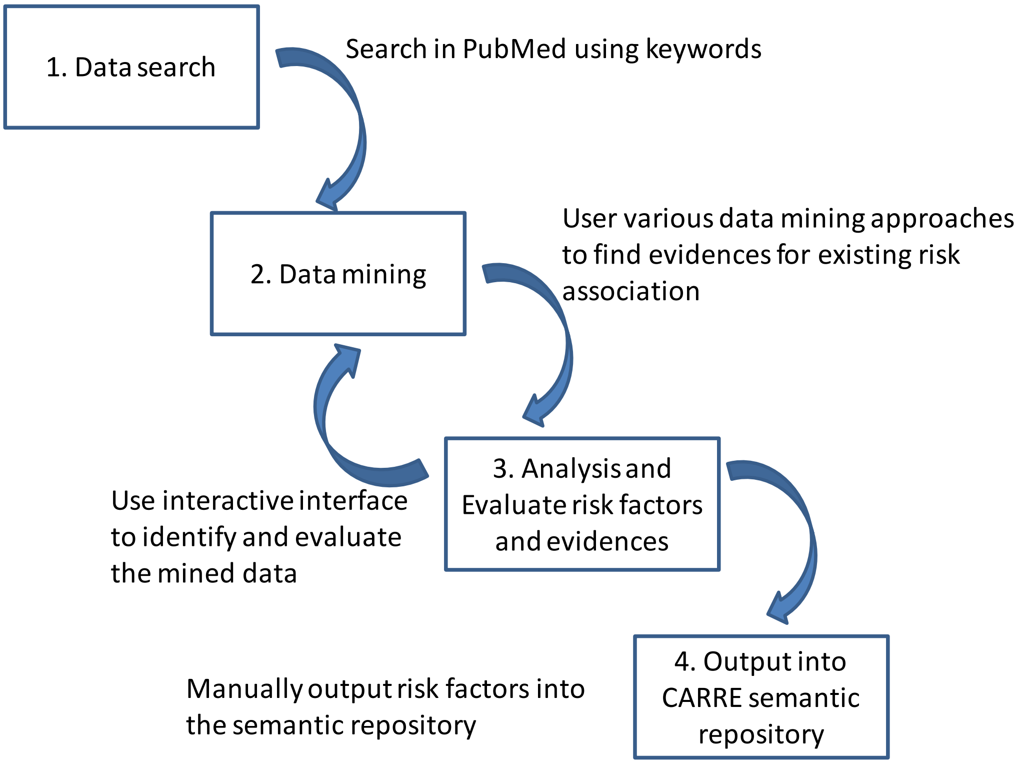

Figure 3. Task breakdown for mining medical evidence.

2.2. Data mining approach

In this section, we will explain the technical background of data mining. The data mining task is based on and extends GATE[3] text engineering framework, shown in Figure 4.

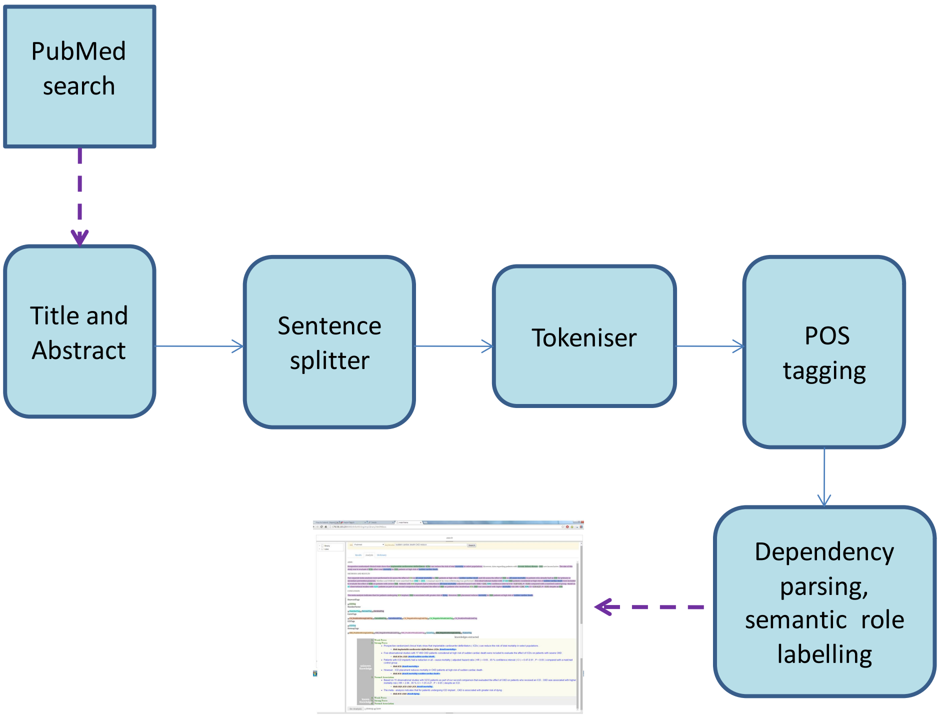

Figure 4. Data mining workflow.

Sentence splitter: is adapted by analyzing the most frequent sentence split patterns/errors.

Tokeniser, uses GATE English language tokeniser, based on a set of regular-expressions and Jape, which provides finite state transduction over annotations based on regular expressions. In the case of English, it breaks a stream of text and gives token types: word, number, symbol, punctuations, and space token.

POS tagging, dependency parsing: it uses Statistical dependency parser, and is Java based, and it uses the CoNLL 2009 data format[4], for example:

- PDEPREL is a syntactic relationship between HEAD (refer to the following table) and this word. It automatically predicted dependency relation to PHEAD, Dependency relation to the head of the current token.
- PPOS is part of speech.

Semantic role labelling: Dependency parsing and semantic role labelling are partly overlapping tasks. It detects semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles.

---

3 https://gate.ac.uk/

4 Jan Hajiˇc, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Ant`onia Mart´, Llu´ıs M`arquez, Adam Meyers, Joakim Nivre, Sebastian Pad´o, Jan ˇSt˘ep´anek, Pavel Straˇn´ak, Mihai Surdeanu, Nianwen Xue, Yi Zhang, 'The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages', Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task, pages 1–18, Boulder, Colorado, June 2009.

The explanations of CoNLL 2009 data dependency format are listed in the following Table 5[5].

Table 5. The semantic role labeling format requires fields.

Field #	Name	Description
1	ID	Token counter, starting at 1 for each new sentence
2	FORM	Form or punctuation symbol (the token; “split” for English)
3	LEMMA	Gold-standard lemma of FORM
4	PLEMMA	Automatically predicted lemma of FORM
5	POS	Gold-standard POS (major POS only)
6	PPOS	Automatically predicted major POS by a language-specific tagger
7	FEAT	Gold-standard morphological features (if applicable)
8	PFEAT	Automatically predicted morphological features (if applicable)
9	HEAD	Gold-standard syntactic head of the current token (ID or 0 if root)
10	PHEAD	Automatically predicted syntactic head
11	DEPREL	Gold-standard syntactic dependency relation (to HEAD)
12	PDEPREL	Automatically predicted dependency relation to PHEAD
13	FILLPRED	Contains ‘Y’ for argument-bearing tokens
14	PRED (sense)	identifier of a semantic “predicate” coming from a current token
15	APREDn	Columns with argument labels for each semantic predicate (in the ID order)

3. Aggregators for scientific data for known risk associations

In this section, we explain our work on searching for medical evidence data for the already known risk factors from available public sources, such as PubMED. The known risk factors are listed in D2.2. As shown in Figure 2, it refers to the functional block ‘identify new evidence’ in *key sentence extractor* and ‘risk analysis’ in *data process*.

3.1. Mining evidences for known risk associations

In CARRE, we adapt a hybrid approach: identify the new evidence using automatic data mining technique; collect and verify detailed evidence data via frontend portal by medical experts to conduct risk analysis on the identified evidence. To achieve this task, we altered and extended the pipeline and showed in Figure 5.

Referring to Figure 3 the data search in PubMED gives us the abstract of a paper that relates to cardiorenal diseases. We then use GATE tool for tokenize and sentence splitter clearNLP[6] is used for POS and tagging.

---

[5] Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task, pages 1–18, Boulder, Colorado, June 2009.

[6] http://www.clearnlp.com/

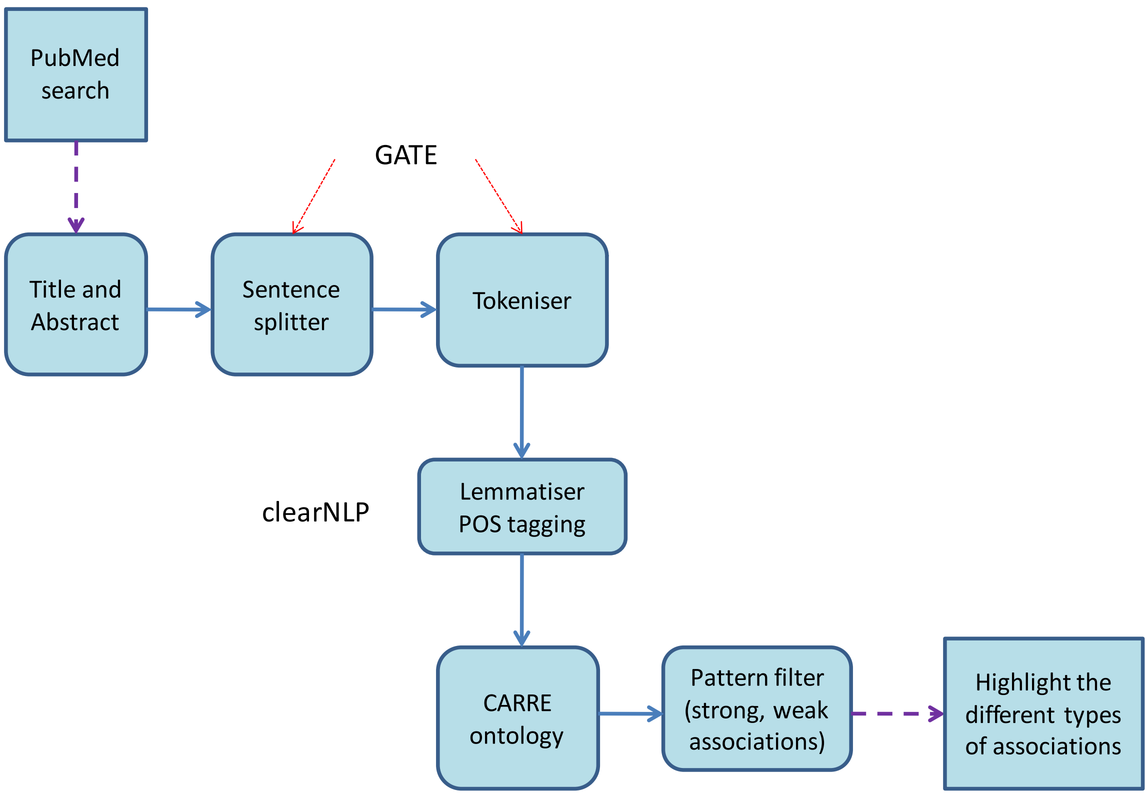

Figure 5. Aggregate known medical evidences.

We use the following example sentence to illustrate the data mining results.

Moderate-severe OSA is associated with type 2 diabetes.

The follow diagram (Figure 6) shows a result of using POS and dependence parsing. As can be seen, the key verb (the ROOT) is 'is', and the dependences between a word to it tokenised head.

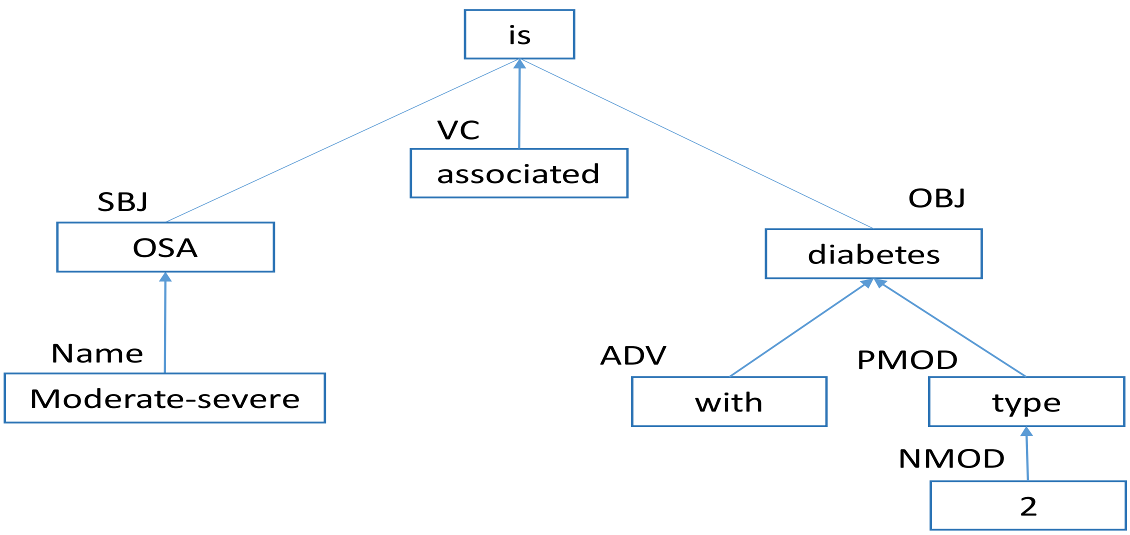

Figure 6. Example of the result of POS dependence parsing.

In addition, we use the following patterns to narrow down the sentences that need to be analysed in order to enhance the accuracy. Based on the above defined risk factors, and the sample papers that we identified by the medical expert, we abstracted the sentences patterns as described in this sections. We define the following:

A: Risk factors (A block)

B: Results (B block)

C: Positive level: much, more, enough, a lot of, lots of, great, numerous, high

Negative level: less, few, low, reduction

Normal level: (the others)

D: Positive description: reduce

Negative description: increase

Normal description: (the others)

E: Certain words related “risk”: risk/risk factor

N: Negative words

Pattern standard[7]: () means optional item and / stands for “or”.

sentence type	(Person)	with	A	have	B
PDEPREL	(SBJ)	CONJ	PMOD	ROOT	OBJ
PPOS8	(NN)	IN	NN(s)	VB(P)	NN(P)

Currently, our CARRE data pattern set has included data type A, B, and N. So the XML data file is required to represent data type C, D, and E. The labels for PDEPREL and PPOS are used to label the words in the sentence, which helps for search action. The data format is explained below:

SBJ: subject

CONJ: conjunction

PMOD: preposition modifiers

ROOT: is the main verb in the sentence

OBJ: Object

NN: noun, singular or mass

IN: preposition/subordinating conjunction

VB: verb, base form

In the Pattern Aa+ and Pattern Ab+, small “a” and “b” express the two patterns are very similar, such as active and passive sentences. “+” represents positive and “-“ means negative.

3.1.1. Relationship A+: Positive Strong Prove

Pattern Aa1+:

(C)	A	D(be/reduce/increase)	E	of/for	B.
(NMOD/SBJ	SBJ/PMOD	ROOT/SUB/VC	OBJ/SBJ/NMOD	NMOD	PMOD/CONJ
(DT)	NN(P)	VB(Z)	NN	IN	NN(S)/VBZ

Sample sentences:

[7] http://www.monlp.com/2011/11/08/part-of-speech-tags/

In summary, this meta-analysis of prospective cohort studies suggests that moderate-severe OSA increases the risk of type 2 diabetes, and the risk of diabetes associated with OSA appears to increase with the severity of OSA.

Findings from several prospective studies indicate that 30 min or more of daily moderate- intensity activity, as recommended in multiple U.S. guidelines, can substantially reduce the risk of type 2 diabetes as compared with being sedentary.

It is commonly recognized that obesity is an established risk factor for type 2 diabetes mellitus.

Recently, the waist-to-height ratio (WHtR) was introduced as the hypothetically best abdominal obesity indicator of risk of type 2 diabetes mellitus because it is reasonable to think that short subjects generally will have more abdominal fat and associated cardiovascular risk factors than will tall subjects under the condition of a similar WC.

The association was partly independent of BMI, suggesting that moderate-intensity physical activity can reduce the risk of type 2 diabetes even in those who do not achieve weight loss.

Pattern Ab1+:

E	of	B	D{be/reduce/increase}	(with)	A
SBJ	NMOD	NMOD/PMOD	ROOT	(ADV)	NMOD/PMOD
NN	IN	NN	VBN	(IN)	NN

Sample sentences:

Relative risk of myocardial infarction increased with tobacco consumption in both men and women and was higher in inhalers than in non-inhalers.

Pattern A2+:

Person	with	A	have	B
SBJ	CONJ	PMOD	ROOT	OBJ
NN	IN	NN(s)	VB(P)	NN(P)

Sample sentences:

Approximately 5% of adults younger than 52 years and without diabetes, hypertension, or obesity have CKD, compared with 68% older than 81 years.

Pattern A3+:

E	of	B	be	number	for	A
SBJ/ APPO	NMOD	PMOD	ROOT	NMOD	NMOD	PMOD
NNP	IN	NN(S)	VB(P/D)	CD	IN	NN

Sample sentences:

The summary RR of type 2 diabetes was 0.69 (95% CI 0.58–0.83) for regular participation in physical activity of moderate intensity as compared with being sedentary. RRS (relative risks)

The summary RR of type 2 diabetes without BMI adjustment was 0.69 (95% CI 0.58 – 0.83) for the highest as compared with the lowest category of moderate-intensity physical activity.

Pattern Aa4+:

(Person)	A	have	(number)	C	E	of	B
(SBJ)	SBJ/PMOD/NMOD	ROOT/SUB	(NMOD)	NMOD	SBJ/OBJ/PMOD	NMOD	PMOD/OBJ
(DT)	NN(S)	VBD/VBP	CD	JJR/VBD	NN	IN	NN(S)/VBZ

Sample sentences:

If underweight subjects had a higher risk of diabetes than those with normal weight across various study populations, the risk of diabetes for incremental increases in obesity could be underestimated.

The meta-analysis suggests that ex-smokers have around a 50% increased risk of suffering a stroke before the age of 75.

In a previous study we showed that female smokers have about a 50% higher relative risk of dying from vascular disease.

Female smokers have a higher relative risk of myocardial infarction than male smokers, even after adjustment for major cardiovascular risk factors.

Pattern Ab4+:

B	have	E	for	A
SBJ	ROOT	PMOD	NMOD	PMOD
NN	VBP	NN	IN	NN

Sample sentences:

Coronary heart diseases have a similar pattern of falling relative risk for smoking with age.

Pattern A5+:

C	in	B	E	for	A
SBJ	LOC	NMOD	PMOD	NMOD	PMOD
NN	IN	NNS	NN	IN	NN

Sample sentences:

After adjustment for BMI, the reduction in diabetes risk remained substantial (17%) for both regular moderately intense activity and walking.

Pattern A6+:

….	A	for	C	of	B	E
	PMOD	ADV	PMOD	NMOD	PMOD	NMOD
	NN	IN	NN	IN	NNS	NN

Sample sentences:

Furthermore, the included studies mostly focused on leisure time physical activity, but commuting and occupational activities can also contribute importantly to the accumulation of moderately intense physical activity for the reduction of diabetes risk.

3.1.2. Relationship B+: Positive Weak Prove

Referring to all patterns of Relationship A+, to build patterns of Relationship B+, number of modal verbs (M) such as “may”, “might”, and “seem to” are introduced to sentences to support the relative verbs (same level or up level to the relative structures) .

For example:

(C)	A	M{may/might/seem to}	D	E	of	B
(NMOD)	SBJ	ROOT/SUB	VC/IM	OBJ/SBJ/NMOD	NMOD	PMOD
(DT)	NN(P)	MD	VB(Z)	NN	IN	NN(S)/VBZ

Sample sentences:

These findings support the hypothesis that OSA may be an independent risk factor for the development of diabetes.

Coronary heart disease seems to have a similar pattern of falling relative risk for smoking with age.

3.1.3. Relationship C+: Positive Normal Association

Pattern C1+:

A/B	(be)	associated/anticipated/related	with/to	B/A
SBJ/PMOD/NMOD/NAME	ROOT/SUB	VC/APPO	ADV/AMOD	OBJ/NMOD/PMOD/CORRD
NN(P)	VBD/VBZ	VBN	IN/TO	NN(S)

Sample sentences:

This meta-analysis indicates that moderate-severe OSA is associated with an increased risk of type 2 diabetes, and this appears to be an independent risk factor for the development of diabetes.

Second, in order to assess whether the severity of OSA was associated with type 2 diabetes, the corresponding pooled risk estimates were respectively evaluated according to the severity of OSA.

Cardiovascular disease mortality was, as anticipated, associated with the full range of risk factors under study, including raised blood pressure, smoking, diabetes, physical inactivity.

In the present individual participant meta-analysis, there was limited evidence that cardiovascular disease risk factors were related to dementia death.

Taking these results together, there was limited evidence that these CVD risk factors were related to the occurrence of dementia death in the current study.

The present meta-analysis indicated that WHtR and WC were more strongly associated with the development of diabetes than was BMI or WHR.

Stroke should therefore be added to the list of diseases related to smoking.

The risks of stroke associated with smoking are apparently present in all age groups but are far greater in younger people.

There has been increasing recognition that obstructive sleep apnoea (OSA) is associated with incident type 2 diabetes.

Pattern C2+:

an/the	association/link	between/of	A/B	and/with	B/A
NMOD	OBJ/PMOD	NMOD	PMOD/NMOD	COORD/NMOD	SBJ/PMOD/NMOD
DT	NN	IN	NN	CC/IN	NN

Sample sentences:

Numerous studies have identified an association between OSA and type 2 diabetes.

The aim of this study was to assess the association between the severity of OSA and the risk of type 2 diabetes by performing a meta-analysis of all available prospective cohort studies.

The Wisconsin study cross-sectional analysis provided evidence of a link between OSA and the development of diabetes; however, in the longitudinal analysis, this association was not significant after adjustment for age, gender and waist circumference.

In addition, the Busselton study identified a significant independent association between moderate-severe OSA and the incidence of diabetes.

Although there was no significant association between mild OSA and increased risk of diabetes in either the Busselton or Wisconsin cohorts, the increased odds ratios were in keeping with an increasing risk of diabetes as the severity of OSA increased.

Therefore, we evaluated the association between OSA and the risk of type 2 diabetes by performing a meta-analysis of prospective cohort studies.

First, the magnitude of the association between OSA and the risk of type 2 diabetes was estimated.

In this article, we systematically review the epidemiological evidence on the association between physical activity of moderate intensity and risk of type 2 diabetes.

The BMI-unadjusted association between moderate-intensity physical activity and diabetes risk was significantly stronger for female (RR 0.58 [95% CI 0.51– 0.65]) than for male (0.82 [0.70– 0.96]) cohorts (P_0.04).

When the study targeted a population with a mean BMI of 28 or greater, the associations between obesity indicators and diabetes risk were significantly or borderline-significantly weakened compared with studies in which the mean BMI was less than 28 (P = 0.02 for RRWHtR, P = 0.04 for RRBMI, P = 0.03 for RRWC, P = 0.11 for RRWHR).

The association of stroke with cigarette smoking seemed to remain after adjustment for alcohol consumption; indeed, adjustment in these studies tended not to reduce the relative risk, which suggests that alcohol consumption is not an important confounding variable.

Pattern C3+:

A/B	have	association/link	with	B/A
SBJ/PMOD	SUB	OBJ	NMOD	PMOD/NMOD
NN(P)	VBD	NNS	IN	NNS

Sample sentences:

This finding is inconsistent with a previous meta-analysis that indicated that BMI, WC, and WHR had similar associations with incident diabetes.

3.2. Negative Relationships

There are five ways to change positive sentences to the negative sentences.

1. verb + not
2. Adding negative adjectives, such as no, inverse, less, and poor etc. for keywords (A, B or association etc.)
3. Adding negative adverbs for the relative verbs, such as seldom, hardly, scarcely, and barely
4. Adding negative prepositions or conjunctions, such as without, except etc.
5. Adding negative nouns or adjectives for a sentence or structure, such as impossible, no way, nothing, and none etc.

Sample sentences:

A randomized trial of moderate-intensity physical activity in individuals with a family history of diabetes did not find a significant reduction in incidence of type 2 diabetes after 2 years, but compliance with the program was poor and the number of participants small (n _ 37 in the exercise program).

The Egger and Begg tests provided no evidence for publication bias for the BMI unadjusted (P _ 0.63 and P _ 0.84, respectively) and BMI-adjusted (P _ 0.83 and P _ 0.40) association between moderately intense physical activity and risk of type 2 diabetes.

In our metaanalysis of 10 prospective cohort studies, a substantial inverse association was observed between physical activity of moderate intensity and risk of type 2 diabetes.

No significant association between moderate intensity physical activity and type 2 diabetes was observed in the two studies that reported results for other ethnic groups, but this may have been due to the “light” definition of activity and limited statistical power as a result of lower numbers for nonwhites.

Results from cross-sectional studies were generally consistent with an inverse association between moderately intense physical activity and type 2 diabetes. We found a significant inverse association between moderately intense physical activity and type 2 diabetes that persisted after adjustment for BMI. Third, publication bias is inevitable under the condition that the association between WHtR and diabetes risk is not commonly recognized.

3.3. Implementation

The implementation uses GATE tools and clearNLP. Both are open source tools. In Figure 7 shows the pipeline analysis with data input and output.

```
1.  pipeline analysis input and output
A.  --------------main invoke--------------
    public void doGet(HttpServletRequest request, HttpServletResponse response)
    throws IOException, ServletException
    {
        //--------------receive parameters from user interface-----------
        String resId,analysisKeywords,dbType,strProject
        //doc id from resources(Pubmed or IEEE)
        //input by user to generate a keywords rule file for Gate
        //dbType is pubmed or IEEE
        //user selected dics from interface
        //when execute rule which project belongs
        //-------------

        String uuid = CommonMethod.UuidGenerator();
        //create original file as url
        boolean flag = GateExtractor.createTaggedFile(resId,uuid);
        String sUrl = null;
        if(flag){
            sUrl = GateExtractor.getTmpFileURL(uuid);//taged file url
        }
        //generate keywords rule
        JapeMain obj = new JapeMain(analysisKeywords,"userName","nodeName");
        String mainFileName = obj.createMainFile();//get main file formatted
        //-------------------init gate processing pipe line
        GateInstance.getInstance();
        //------read pattern file first,init rules
        CarreRule.getInstance();
        //----------init PR
        PipelineFactory.initProcessingResouces();
        //----------init dics PR------
        GateExtractor extObj = new GateExtractor( mainFileName,analysisKeywords,mode);//0 for stand alone
        extObj.initSerialAnnieNew(dics);
        //---------------------------create corpus--------
        Corpus corpus = Factory.newCorpus("GateParser corpus");
        //-------create a file----------
        flag = extObj.createCorpusDocument(corpus,dbType, resId,uuid) ;//dbType temp GateExtract.FileNameEnum{Pubmed,IEEE}
        // tell the pipeline about the corpus and run it
        if(flag){
            extObj.setCorpus(corpus);
            extObj.execute();

            String resultFileURL = extObj.processDoc(corpus,project,0);

        }
    }
//------------------------explanition---------
resultFileURL: http://10.1.62.239:5984/db-test11/471eae90_027a_4e01_b36d_b87350fd32f0/ResultFile.html
CarreRule.getInstance():init pattern file and load models
initProcessingResouces: init basic gate process resources
initSerialAnnieNew: init dictionaries and ontology pipe lines for all those dics
```

Figure 7. The main pipeline code.

The CARRE rules mentioned in Figure 7 are based on the patterns in section 3.1 and 3.2. The rule class format explains the matching process of the searched sentences against the above-mentioned patterns. The file of data format is as follows:

List: //different dictionaries (word sets)

A //Representing letter: more //including the words (1 // present the level 0: normal 1: increase 2: reduce), much (1), less (2)

B: risk, risk factor // “no ()” represents these words without a level description

Pattern: // different patterns

Class: Strong Prove // the property of a class

Patternname: Aa1+ // the name of this pattern

Words // rule of words :Ri;C,D;;M;,E;;associate;,of/for,Re

POS // rule of POS:

Structure // rule of structure:

In the rule, there are the concept of the cell using “,” to separate. In each cell, there are four items separating by “;”, namely necessary (Npa), optional (Opa), avoiding (Apa), ignoring (Ipa) items.

Figure 8 shows resource processing where you can see calls to the pipeline components in Figure 5.

```java
B.---------------- initProcessingResouces ------------
    //be invoked before AssembleProcessingResource or in servlet
    public static boolean initProcessingResouces(EnumMode mode){
        boolean rtn = false;
        try {
            //------------basic pr begin
            getDocumentResetPR();
            getTokenizerPR();
            getSentenceSplitterPR();
            getPosPR();
            getMorphPR();
            //-------------basic pr finished
            CommonMethod.screenPrint("in PipelineFactory basic pr finished");
            initDictionaries(mode);
        } catch (ResourceInstantiationException e) {
            // TODO Auto-generated catch block
            CommonMethod.screenPrint("in PipelineFactory.initProcessingResource:"+e.getMessage());
            e.printStackTrace();
        }
        rtn = true;
        return rtn;
    }
```

Figure 8. Initial resource processing.

3.4. The frontend user interface

The interface, as shown in Figure 9, provides highlighted found risk associations, for experts to check and further investigate the risk associations. The service can be reached at http://176.58.103.20:8080/mha. The web interface is used for experts to check and confirm the founded evidence and new risk associations.

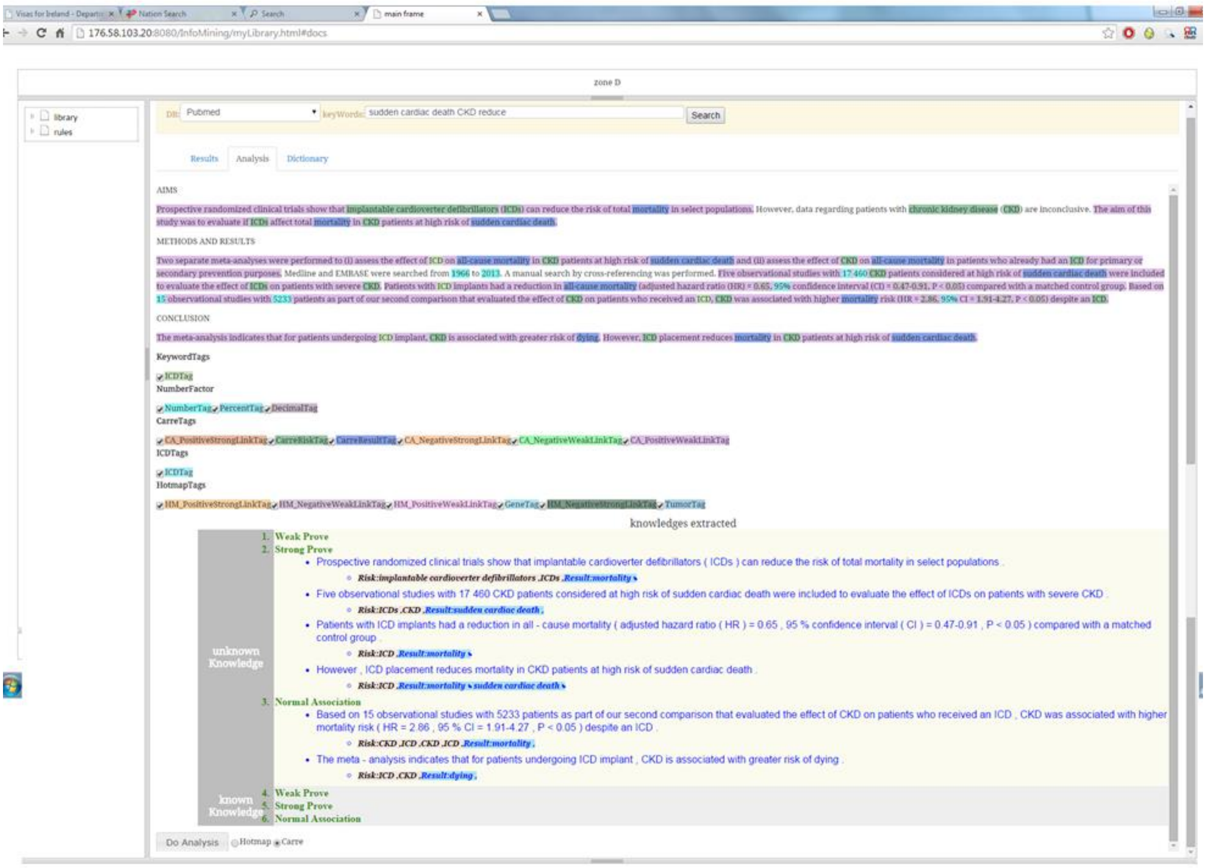

Figure 9. Screenshot of the user interface showing system output.

4. Aggregators for unknown risk associations

The ultimate goal of the scientific literature aggregator is to search for unknown medical evidence data stored in openly available public sources, such as the PubMED abstract indexing service and database. The service acts as a web crawler and searches the new risk factors at the backend. The collected literature for both supportive information and new risk factors will be evaluated by medical experts and, if approved, new information on risk factor descriptions will be stored in the CARRE public semantic repository.

4.1. The pipeline for finding new risk associations

With regard to the new risk associations, we need to find unknown risks for the cardiorenal diseases, but the risk could be properly defined medical terms or it may be „normal" words, such as *running for more than an hour*. Therefore, in addition to the components we used in section 3, we use topic modelling (refer to Figure 2) as assistant in mining the new risk associations. A topic is represented by a fixed number of linguistic elements, such as action type of words, called VAs and VACs. We follow a revised pipeline as shown in 10. Semantic role labels from clearNLP are added to the data mining process. Together with tokeniser, sentence splitter, POS, dependency parsing, sentences are analysed regardless of the medical knowledge, or any other domain knowledge.

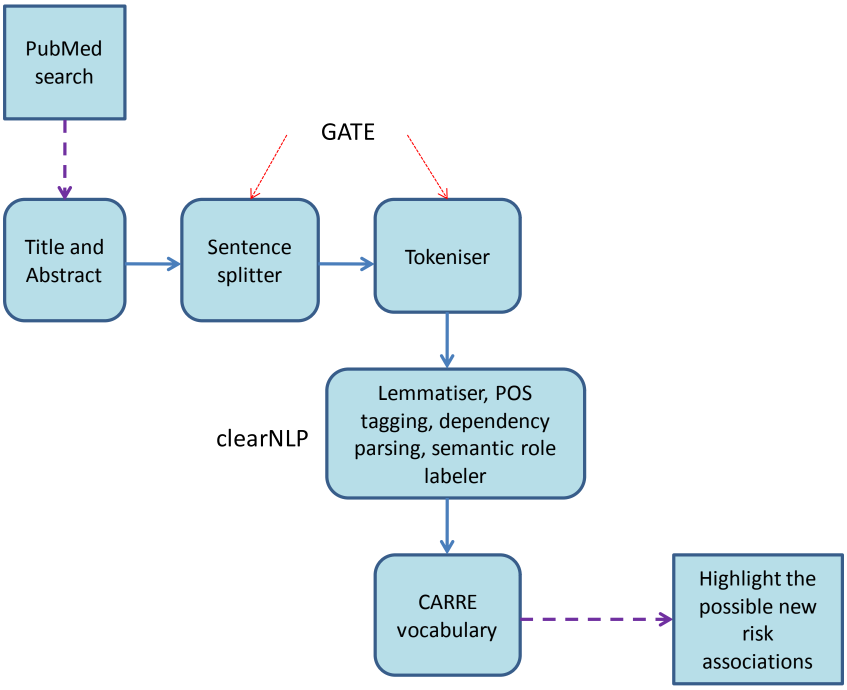

Figure 10. Medical scientific literature aggregator workflow.

Let us use an example to explain the approach. The following output is from MATE[8] tool just for the purpose of explanation.

In summary, this meta-analysis of prospective cohort studies suggests that moderate-severe OSA increases the risk of type 2 diabetes.

Figure 11 shows the output as a result of dependency parsing and semantic role labelling.

[8] https://code.google.com/p/mate-tools/

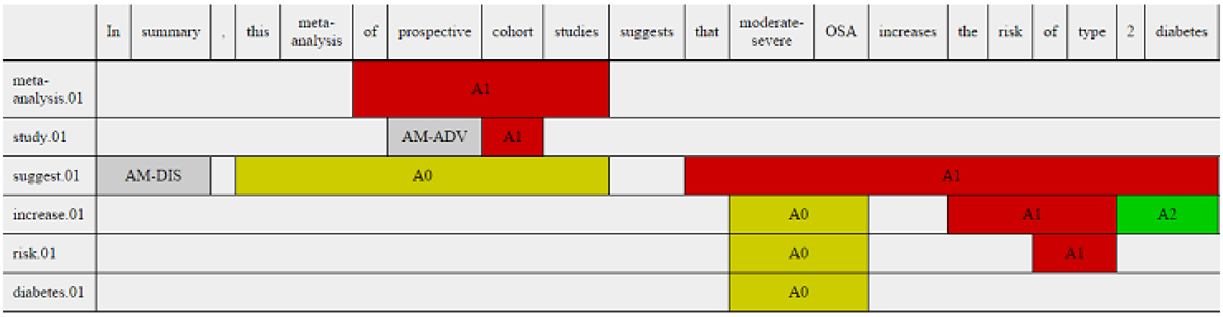

Figure 11. Output of dependency parsing and semantic role labelling.

As can be seen from the above example, on the left hand side, MATE lists al nouns and verbs. The identified A0 and A1 have already picked up the main risk associations. Our approach starts from the finest grind A0 and A1. In this case, they are moderate-severe OSA, and type 2 diabetes. To make sure they are the risk associations that we need for CARRE project, A0 and A1 will be further analysed using domain knowledge. Medical partner VULSK had reduced MeSH[9] vocabulary to a subset including only terms related to cardiorenal disease. Since the risk result can be found in either the subject or object, we check both A0 and A1 for the result; if the result appears in A0, then we check the corresponding A1 for the risk, and vice versa. The corresponding A0 and A1 means that they are related to the same verb, refer to Figure 11, both *moderate-severe OSA* and *type 2 diabetes* are in subordinate clause and *increase* is the verb. In either A0 or A1, apply the CARRE vocabulary to pick up the results and then the risk. In the above example, *diabetes* as a disease should be picked up and hence the relevant *OSA*. Figure 12 shows the sample search results.

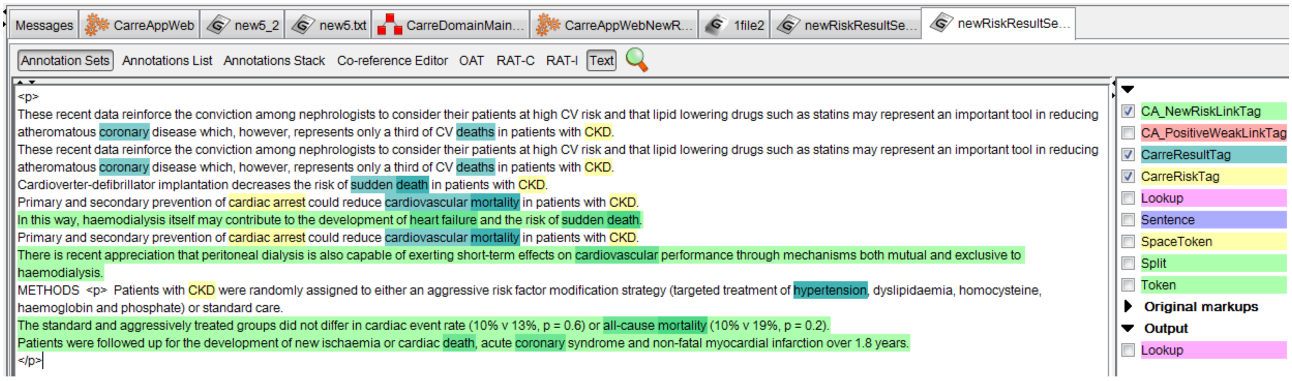

Figure 12. Screenshot of finding new risk factors.

9 Medical Subject Searching controlled vocabulary, http://www.nlm.nih.gov/mesh/

4.2. Implementation

Figure 13 shows how to use clearNLP to identify data dependence in a sentence. As a result, A0 and A1 can be identified. Figure 14 shows the way to analyse sentences that related to CARRE risk associations. The CARRE vocabulary set is used in extracting the sentences for further analysis.

```
/*
 *
SentenceRelationModel includes these informations:
ID:  FORM       LEMMA     POS    FEATS        HEAD   DEPREL    SHEADS
1    I          i          PRP    _              3   nsubj     3:A0;5:A0
2    'd         would      MD     _              3   aux       3:AM-MOD
3    like       like       VB     pb=like.02     0   root
4    to         to         TO     _              5   aux
5    meet       meet       VB     pb=meet.01     3   xcomp     3:A1
6    Mr.        mr.        NNP    _              7   nn
7    Choi       choi       NNP    _              5   dobj      5:A1
8    .          .          .      _              3   punct
* */
public static boolean Extraxtinf(SentenceRelationModel aSentenceModel)
        {
            ArrayList<SentenceWordModel> Senlist=aSentenceModel.getTheSentence();//get word list
            String[] StrrevHead=new String[aSentenceModel.getTheSentence().size()+1];//word ArrayList
            String[] StrpaID=new String[aSentenceModel.getTheSentence().size()+1];////word ArrayList
            String rawSentence = aSentenceModel.getRawSentence();//original Sentence

            ArgslistModel Argslist;//=new ArgslistModel();
            ArrayList<ArgslistModel> ArraArgslist=new ArrayList<ArgslistModel>();
            ArgsModel ArgsModel=new ArgsModel();

            //fill in ArraArgslist word by word deal with pHead
            for (SentenceWordModel aWord:aSentenceModel.getTheSentence()) //loop word ArrayList
            {
                    //... if a word contains Args information add it in a list
                    ArraArgslist.add(Argslist); //add argslist for a word
            }//exp sentence: we also introduce a novel method of handle collisions around crease.

            for (SentenceWordModel aWord:aSentenceModel.getTheSentence())//loop sentence again deal with args
            {

                    for (int i=0;i<strshead.length;i++)//loop heads//exp:3:A0;5:A0 //we:3:A0=PAG
                    {
                        //... find start/end word for a Arg
                    }//end of for

                }

            }
            aSentenceModel.setArgslist(ArraArgslist);
            return true;
        }
```

Figure 13. Sentence analysis.

```
//-------------carreDomainSentenceExtractNew.jape----------------
Phase: DOMAINANY
Input: Lookup Sentence
Options: control = all
Rule: Matchingdom
({Sentence}):sent
-->
:sent{
 AnnotationSet lookups = inputAS.get("Lookup").getContained(sentAnnots.firstNode().getOffset(),sentAnnots.lastNode().getOffset());
 String strURI="http://www.DIV.org/Carredom/1#";
 String strrisk=strURI+"Risk";
 String strresult=strURI+"Result";
 String strcause=strURI+"Causes";
 String strcausedby=strURI+"Caused_by";
 String strnegative=strURI+"Negativevoc";
 long intresult=0; intrisk=0;intneg=0;
 for (Annotation lk:lookups)
 {
     if ("instance".equals((String)lk.getFeatures().get("type")))
     {
           String strinstance=(String)lk.getFeatures().get("URI");
           OInstance ainstance = ontology.getOInstance(ontology.createOURI(strinstance));
           if(ainstance!=null)
           {
               if (strrisk.equals(lk.getFeatures().get("classURI")))//a risk
               {
                    intrisk++;
                }else if (strresult.equals(lk.getFeatures().get("classURI")))//a result
               {
                    intresult++;
                }else if(strnegative.equals(lk.getFeatures().get("classURI")))//a negative
               {
                    intneg++;
                }
          }
    }
    if (intresult>0 && intrisk==0){
        inttype=5;
     }
     sentAnnots.iterator().next().getFeatures().put("intresult",intresult);
     sentAnnots.iterator().next().getFeatures().put("intrisk",intrisk);
     sentAnnots.iterator().next().getFeatures().put("intneg",intneg);
     sentAnnots.iterator().next().getFeatures().put("inttype",inttype);
}
```

Figure 14. Sentence extractor.

4.3. Code metrics

The project used some open source software, such as LGPL for Gate, Apache License for clearNLP, Maven and MIT license for front end Jquery. Source code for both frontend and backend are hosted on https://bitbucket.org/weihuiBeds/carre-text. The current implementation utilises Apache Maven for software project management and comprehension tool.

Code quality analysis has been conducted on the Java file. We use eclipse plugin PMD for Code check, and eclipse plugin CodePro AnalytiX for code metrics. The results are shown in Figure 15. Apart from the sample charts, the detailed statistics, such as *lines of code*, are shown on the left-hand side in the diagram.

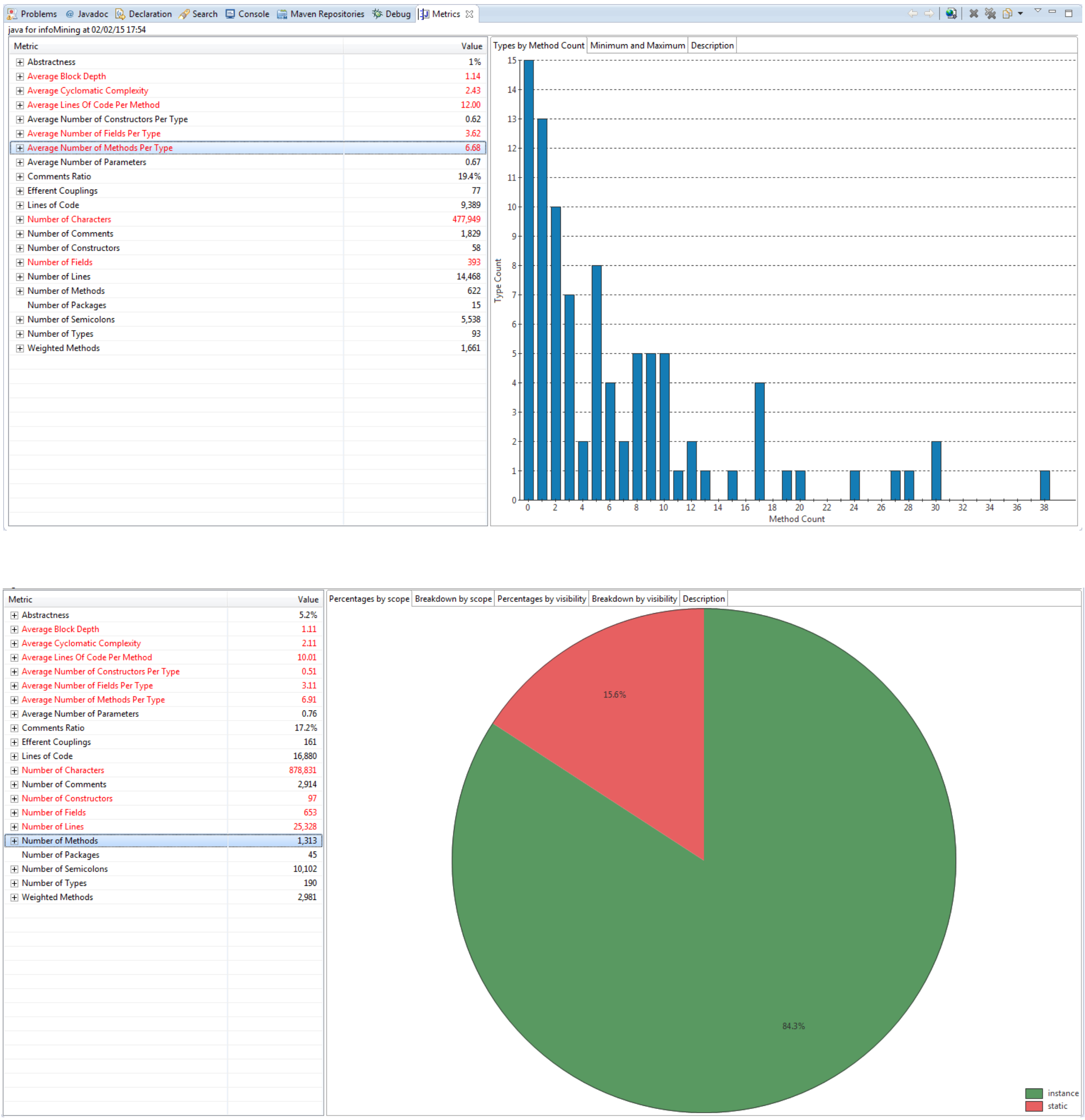

Figure 15. Code metric results.

4.4. Discussion

Identifying risk associations should be conducted in a semi-automatic way. For example, as shown in the following example, the key words can be highlighted as follows based on the approaches we described in section 4. However, this sentence is not the risk association we are looking for and should be removed.

Annual cardiovascular mortality in patients with chronic kidney disease (CKD) is much higher than in the general population.

In addition, in the implementation, we cannot automatically and accurately collect different types of numbers as support evidence, therefore the user intervention is needed to identify and collect for these types of data.

The user interface serves the purpose.

Refer to Figure 2, general statistics can be applied to get sentences patterns of describing risk factors. In addition, the Research Object System (ROS) is a generic data-mining component that analyses sentences and separating the structures from sentences. It could potentially further improve the accuracy of the mining results.

5. Risk Model Semantic Data Entry System

This section explains the implementation for step 4 in Figure 3. The Risk Model Semantic Data Entry system - **RMSDE** - was initially developed in order to capture the risk associations identified in D.2.2. Hence, our initial objective was to provide a computer-human interface for our CARRE medical experts and avoid dealing with files that reside in multiple machines.

In order to speed up the development of the desired application, the OU team decided to use Drupal[10] as a Web Application Development Environment. Drupal, which is also considered a Content Management System constitutes a flexible and generic environment that allows developers to customise the way content is created, viewed and consumed. Drupal supports the above customisation with a number of different modules that are developed, extended and supported by the open source community. As of January 2015, there are more than 10,000 open source modules that address many different needs[11].

In the remainder of the section, we discuss the key features that characterise the above system, emphasising on its technical aspects.

5.1. Custom Content types and rich web forms

As expected, the content that is designed to describe risk associations introduces a number of CARRE-related concepts that are not supported natively by Drupal. In order to overcome this limitation, we used a native Drupal module that allows the custom creation of content types. The result of this adoption is that we ended up creating the following content types:

- *Citation*
- *Observable*
- *Risk element*
- *Risk evidence*
- *Risk factor*

The above content types correspond to the core CARRE classes identified in the CARRE vocabulary as introduced in D.2.4. Each class is comprised of a number of primitive data types (e.g. free-text, integer,

[10] https://www.drupal.org

[11] https://www.drupal.org/project/project_module and browse modules for Drupal Version 7.

selection list, Boolean etc.). For example, a *Risk Element* contains a checkbox for capturing the Boolean value of whether the risk element is modifiable or not (attribute *Modifiable*).

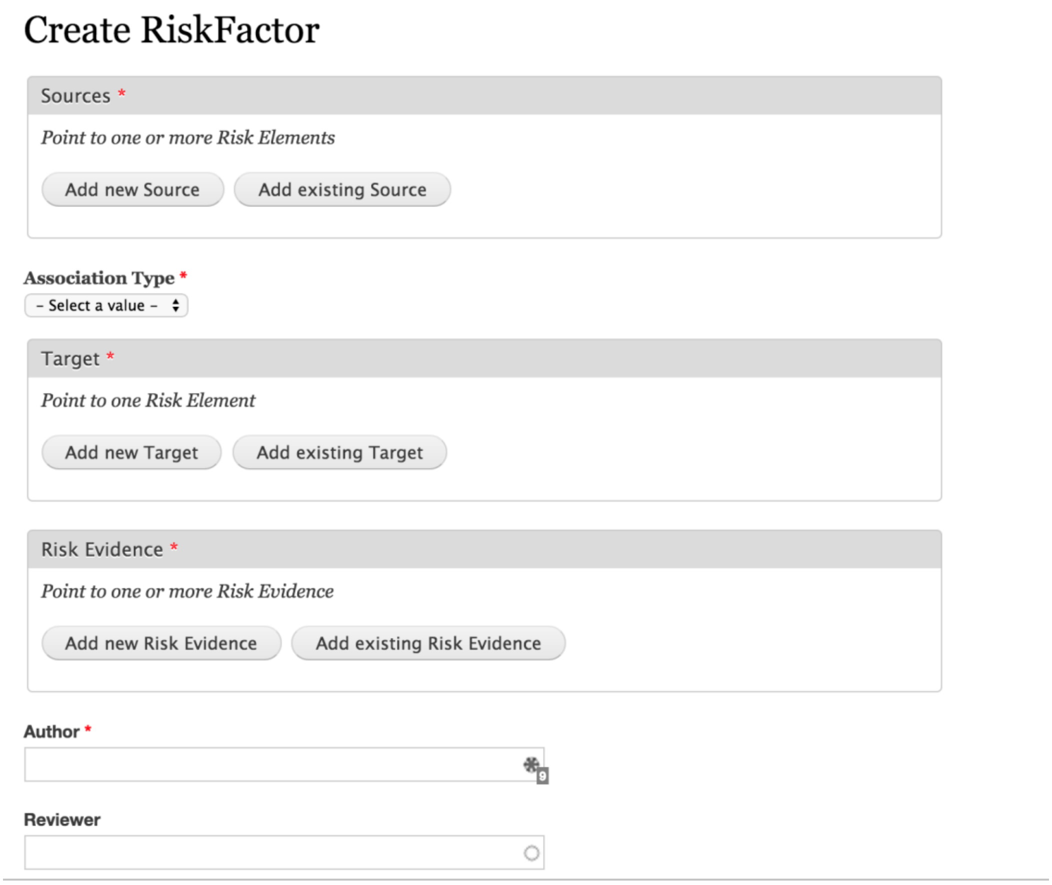

Figure 16. Screenshot of Risk Factor creation form.

After all content types are created, we assigned Create-Read-Update-Delete (CRUD) permissions to authenticated users in order to allow them to proceed with the data entry process.

Furthermore, some content types function as placeholders for other CARRE content types and are more complex objects. An example of this object is an instance of a *Risk Factor*. An instance of a Risk Factor contains one or more Risk Elements as *Source*, one or more Risk Elements as *Target*, while it also contains at least one *Risk Evidence*. In order to allow our end-users to create such complex objects, we used a dedicated module called “Inline Entity Form”[12] which allows the creation and reuse of Drupal content-objects. A screenshot of a form that uses the above functionality is shown in Figure 16, which illustrates the *Risk Factor* creation form.

5.2. Connection to external repositories

Part of the data anticipated in the RMSDE system already resides in 3rd party, external repositories. An example of this data is a PubMED publication. In order to accommodate the efficient reuse of this data, we have developed a custom module (called PubMED) which makes use of the PubMED API[13]. Our module is accessing the PubMED API, looks up the citation metadata and automatically inserts them into the RMSDE system. Figure 17 shows a screenshot of how a CARRE user may insert one citation using the PubMED identifier. Figure 18 shows a view of RMSDE that lists PubMED citations inserted.

[12] https://www.drupal.org/project/inline_entity_form

[13] http://www.ncbi.nlm.nih.gov/books/NBK25500/

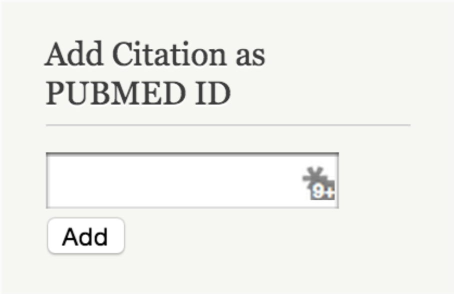

Figure 17. A screenshot with the textbox that allows the insertion of PubMED publications using PubMED ID.

Citations

Title	Authors	Year	PMID	Issue	Journal	Source type	Volume	Delete link
Chronic kidney disease after acute kidney injury: a systematic review and meta-analysis.	Coca SG, Singanamala S, Parikh CR	2012	22113526	5	Kidney international	0	81	delete
Predictors of new-onset heart failure: differences in preserved versus reduced ejection fraction.	Ho JE, Lyass A, Lee DS, Vasan RS, Kannel WB, Larson MG, Levy D	2013	23271790	2	Circulation. Heart failure	0	6	delete

Figure 18. Screenshot showing citations inserted into RMSDE.

In addition to a direct insertion using an already known identifier, CARRE users may search in PubMED repositories, select and insert the publication they desire in one click without having to go to PubMED (see Figure 19).

PUBMED search results

View Edit

- Evaluation of chronic kidney disease in chronic heart failure: From biomarkers to arterial renal resistances.
  Iacoviello M, Leone M, Antoncecchi V, Ciccone MM
  World journal of clinical cases 1(3)
  PMID: 25610846
  insert
- Vitamin D analogues to target residual proteinuria: potential impact on cardiorenal outcomes.
  Humalda JK, Goldsmith DJ, Thadhani R, de Borst MH
  Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association
  PMID: 25609737
  insert
- B-type natriuretic peptide as a predictor of ischemia/reperfusion injury immediately after myocardial reperfusion in patients with ST-segment elevation acute myocardial infarction.
  Arakawa K, Himeno H, Kirigaya J, Otomo F, Matsushita K, Nakahashi H, Shimizu S, Nitta M, Takamizawa T, Yano H, Endo M, Kanna M, Kimura K, Umemura S
  European heart journal. Acute cardiovascular care
  PMID: 25609593
  insert

Figure 19. A screenshot with the search results coming from PubMED. Notice the "insert" link that can be used to automatically fetch all citation metadata.

5.3. RDF and SPARQL endpoint

Drupal[14] by default stores all of its content into MySQL, which is a relational Database. Hence it would require additional effort to transform relational data into the appropriate RDF triples. In order to address this need, we have initially made use of the Semantic Web toolset provided by Drupal's default installation. Drupal allows the annotation of content types' fields with custom RDF predicates. More specifically, we have imported the CARRE vocabulary together with other vocabularies introduced in D.2.4 and used the same terminology to describe the Drupal fields.

Following the above steps, we have installed an extra module that exposes all content through a SPARQL endpoint[15]. The result is that RMSDE exposes all content in a format that allows the automatic migration of its data to the dedicated RDF repository, which is running on a Virtuoso-powered server (see D.2.5 for the overall architecture of CARRE).

5.4. Limitations

While the RMSDE has managed to successfully capture the information for which it was designed, it also suffers from a few limitations. The main limitation is that this system is not tightly connected to the overall CARRE architecture. This limitation stems from the decision to use Drupal, which is tightly coupled with a relational database as a backend (MySQL) and does not allow the seamless integration with the main CARRE repositories (or any other repository). This constraint requires from users to create two user accounts and requires switching between different environments (i.e. RMSDE and remaining of the upcoming CARRE platform). Moreover, the Drupal-powered RMSDE is introducing extra technical effort when considering its integration with the other Aggregators of Medical Evidence discussed in this deliverable. Hence, a new version of RMSDE is currently under development that will address all of the above needs. We intend to present this new version in the coming deliverables.

6. Aggregators for educational data

The aim of the educational resource aggregator is to harvest educational resources from 3rd party repositories, present these to the medical expert for annotation and rating, and output the results of the annotation (together with resource metadata) to the CARRE public RDF repository.

In particular, field survey as presented in CARRE D.2.3 "Data Source Identification and Description" shows that on-line repositories with information for patients are increasing in number and content. In this pilot implementation in CARRE we chose to harvest repositories that are free of charge and provide an API. In order to allow comparative demonstration and appraisal, we chose two different representative repositories: (a) MedLinePlus, an authoritative repository, HONcode[16] compliant and provided by an established scientific body; and (b) Wikipedia, the most popular public encyclopedia, freely developed by crowd contributions. Detailed descriptions of these repositories are given in D.2.3.

The purpose of the aggregators developed here is to harvest metadata about related educational content. These metadata are further enriched by semantic interlinking using controlled vocabularies available in Bioportal as well as semantic tags provided by DBpedia.

BioPortal[17] is an open repository of biomedical ontologies that provides access via Web services and Web browsers to ontologies developed in various formats including OWL, RDF, OBO format and Protégé frames. Amongst the more than 420 ontologies included, there are prominent medical ontologies such as SNOMED-CT (Systematized Nomenclature of Medicine – Clinical Terms), ICD9/10 (International Statistical Classification Diseases and Related Health Problems), Body System (body system terms used in ICD11), MeSH (Medical Subject Headings), NCI (Meta)Thesaurus, Galen (the high level ontology for the medical domain).

---

14 https://www.drupal.org/project/sparql

15 The SPARQL endpoint for the RMSDE is https://carre.kmi.open.ac.uk/sparql

16 http://www.hon.ch/HONcode/Patients/Visitor/visitor.html

17 http://bioportal.bioontology.org

The educational resource aggregator uses the NCBO's RESTful Web services programming interface to access and incorporate terms and concepts from the more than 420 ontologies provided to this day, corresponding to more than 6 million medical and life sciences terms. This way the aggregator can help the user annotate an educational resource with suggested standardized terms and concepts from a variety of ontologies, enriching the RDF output with dereferencable standardized terms as values for the various fields, e.g. keywords, discipline, specialty, etc.

6.1. Architecture

The overall aggregator architecture is shown in Figure 20. The aggregator has a backend and a frontend.

The main parts of the backend are the Resource Retriever, the Resource Rating and the Resource Metadata Processing. In short, the Resource Retriever accepts CARRE concept terms from the CARRE public RDF repository and uses them to formulate queries to external 3rd party educational resource repositories. The results of this search are parsed to extract metadata. Then the retrieved results and metadata are displayed to the expert user for rating and annotation (via the aggregator front end). Rating involves expert user opinion and annotation as well as subjective measures calculating by the Educational Object Rating Module. Expert rating will involve assessment of content-keyword relevance, content accuracy and depth of coverage, while the automatic systems rating will involve Readability Test based on the Flesch-Kincaid algorithm[18], and rating based on the latest modified version of the article and number of revisions.

Finally, Resource Metadata Processing involves metadata enrichment via semantic web sources (such as NCBO Bioportal medical ontologies and DBpedia) and mapping to the CARRE RDF schema so that metadata can be pushed to the CARRE public RDF repository.

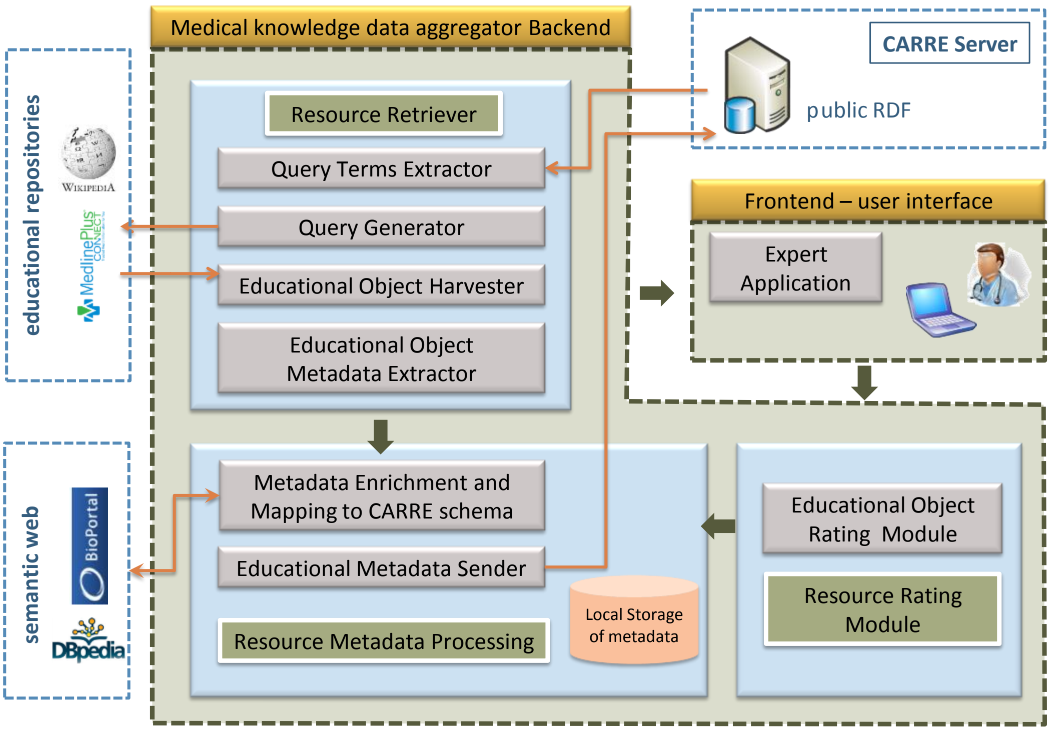

Figure 20. Architecture of the educational resource aggregator.

[18] Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS (1975). "Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel]". Research Branch Report 8-75, Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

The *Query Terms Extractor* makes a request to the public RDF server and gets a response with all the relevant CARRE terms (observables, risk elements, risk factors), which then are sent to the *Query Generator* and assist the user in constructing the educational material search query.

Then, the search query is forwarded to the external educational repositories and accepts as a response the list of educational resources that are subsequently forwarded to the *Educational Object Harvester*. The *Educational Object Harvester* applies specific filters to each repository in order to eliminate irrelevant material that does not meet the requirements of CARRE. The refined educational material will be classified as at least one medical database identifier such as MESH, ICD, UMLS, or other relevant controlled vocabulary identifier included in the external database response.

Next step in the processing unit is the enrichment of metadata which provides extra information for the education resource and initial metadata provided by the *Educational Object Harvester*. Here the backend component searches the semantic repositories of DBpedia and Bioportal in order to extract:

- supplementary identifiers;
- alternative labels;
- relevant concepts;
- further classification and categorization; and
- languages which the resource is available in.

After all data is collected an RDF schema is used for integrating all metadata information into CARRE public RDF and made available to the LOD [19] cloud.

6.1.1. Educational resource description

Each educational resource is described by a set of attributes as harvested by the external repository and further enriched by the aggregator. These attributes are grouped in three conceptual categories as described below.

The first part is essentially typical properties of a digital document, referring to provenance:

- URI
- Date published or created
- Title
- Publisher
- Copyrights
- Language

The second part consists of statistical document description:

- Date Modified
- Number of revisions
- Views
- Reviews
- Words count
- Social media mentions (likes , tweets)
- Typical viewing duration (in minutes)
- References/Citations
- PageRank score[20]

---

[19] The Linking Open Data cloud : http://lod-cloud.net

[20] The PageRank Citation Ranking : http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

The third part incorporates semantic profiling of the document[21]. Terms and concepts were collected from various projects like LOM [22] and more in depth from specifications of HCLOM, DCMI, mEducator, LRMI which will be described further ahead. The content and context of the document:

- Semantic Density[23]
- Difficulty
- Assessment of Readability[24]
- Typical reading time
- Audience
- Audience educational level
- Depth of coverage
- Validity
- Accuracy[25]
- Controversial content
- Audience
- Search Query

6.1.2. Educational resource rating model

This section describes the criteria used in the process of the expert rating of the resource. The rating model consists of a subset of the main educational description with some context derived from the user interaction process. The final user rating criteria converge into the following:

- Difficulty
- Depth of coverage
- Validity
- Accuracy
- Controversial content
- Relevance

The document model will consist of user-generated content as well as system-auto generated. The attributes that are filled in via automatic calculations within the aggregator are:

- Date Modified
- Assessment of readability
- References/citations
- PageRank score
- Social media mentions (likes , tweets)
- Synonyms
- Relative categories or concepts according to BioPortal and DBpedia

---

21 http://www.sciencedirect.com/science/article/pii/S0360131502000180
22 http://en.wikipedia.org/wiki/Learning_object_metadata
23 http://bioportal.bioontology.org/annotator
24 http://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests
25 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3170066/

– Medical resource identifiers
  - eMedicine[26]
  - ICD10[27]
  - MeSH
  - CUI (NLM UMLS[28])

These attributes are calculated in the backend the moment the user attempts to read and rate a specific article and are finalized before the export to the CARRE repository.

6.2. Implementation

The web application of the educational resource aggregator is accessible at http://edu.carre-project.eu/ and also provided through the public project web site http://www.carre-project.eu/innovation/educational-data-aggregator/.

The technology stack involves JavaScript on both backend and frontend, all software used is licensed with compliance of the Open Source Definition[29]. At the frontend, the AngularJS[30] framework is used for separating the design from logic and providing rich user experience by handling all requests asynchronously. The backend environment integrates a NodeJS[31] application server that handles simultaneous requests to the external educational repositories such as Wikipedia, MedlinePLUS etc. The local storage is implemented by MongoDB[32] , a document-oriented NOSQL[33] database.

The backend module is being implemented as small independent components having in mind the micro-services architecture.

The *Resource Retriever* consists of 2 services that make use of SPARQL protocol in the case of query term extraction from the CARRE server and API requests to each educational repository. Examples of the requests are presented below:

1. SPARQL query to public RDF (http://carre.kmi.open.ac.uk/sparql) , e.g.

```
SELECT DISTINCT * WHERE {
        ?uri http://purl.org/dc/terms/title ?title .
        ?uri http://www.w3.org/1999/02/22-rdf-syntax-ns#type http://rdfs.org/sioc/ns#Item .
        ?uri http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://xmlns.com/foaf/0.1/Document .
} LIMIT 500";
```

2. API request to Wikipedia and fetches , using the nodeMW [34]library

```
var queryParams={
    'action': 'parse',
    'page':title,
```

[26] http://emedicine.medscape.com/
[27] http://apps.who.int/classifications/icd10/browse/2010/en
[28] http://www.nlm.nih.gov/research/umls/
[29] http://opensource.org
[30] https://angularjs.org
[31] http://nodejs.org
[32] https://mongodb.org
[33] http://en.wikipedia.org/wiki/NoSQL
[34] https://github.com/macbre/nodemw

```
        'prop':'displayTitle|text'
    };
    wiki.api.call( queryParams,function(data) {
        return res.status(200).send(data);
    });
```

3. API request to MedlinePLUS , e.g.

http://wsearch.nlm.nih.gov/ws/query?db=healthTopics&term=cancer&rettype=brief

The *Resource Metadata Processing* unit is a combination of 3 services that collect data per article, making multiple SPARQL requests to enrich the data and finally store it as a unique identified resource into the local MongoDB datastore. Then data is transformed into RDF triples in order to be inserted to CARRE educational repository.

1. DBpedia SPARQL requests to their public endpoint (DBpedia.org/sparql) and parse of results for the purpose of supplementary metadata.
2. Bioportal API requests to the recommender tool[35] to provide relative concepts, unique identifiers
3. Each educational resource data is converted to a SPARQL update query using RDF[36] NMP package and published to the CARRE endpoint NMP

The **Resource Rating Module** is an optional process that requires special user privileges. Each user role is assigned to different rating criteria according to user authorization. As of version 0.2, only 2 user roles have been taken into implementation.

The Expert Doctor user role is assigned with the following criteria as seen in Figure 21:

1. Depth of coverage
2. Accuracy
3. Comprehensiveness
4. Educational level
5. Relevancy
6. Validity

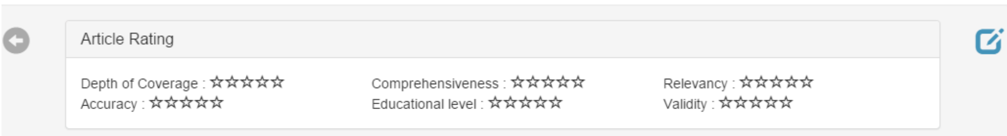

Figure 21. Rating criteria for the Expert Doctor.

Another role authorized for rating is the public User, who can provide simple 5-star rating. It is important to mention that the public user must login with local or social account in order to access the rating component. This is done to secure rating from malicious actions that may lead to rating distortion.

The frontend module is built on top of the AngularJS MVVM framework, a similar approach of the micro-services architecture [37]but on the client side. Using same architecture on both frontend and backend leads to a number of advantages such as better code maintainability because the folder and code structure is similar

35 http://data.bioontology.org/documentation#nav_recommender

36 https://www.npmjs.com/package/rdf

37 http://en.wikipedia.org/wiki/Microservices

to the server, backend agnostic – the frontend communicates with API's , so as long as the API stays the same the application continues to work.

The visible components of the web application are built upon html5 and CSS3 using modern frameworks for consistency and responsiveness like Twitter Bootstrap[38] CSS Framework have been extensively used. Bootstrap is also responsible for mobile/tablet view of the web application.

6.2.1. Deployment specifications

Software deployment is supported for Unix like machines (Linux, Mac) and requires the following libraries to be installed:

- NodeJS application sever
- MongoDB database server
- Git version control system

Next the Educational Aggregator repository[39] should be cloned from Github, all dependencies installed and the build script executed. The commands for the above steps are shown in Figure 22.

deployment setup commands

```
$ git clone https://github.com/telemed-duth/carre-edu.git
$ cd carre-edu
$ npm install && bower install
$ grunt serve
```

Figure 22. Commands for setting up the deployment of the educational resource aggregator.

6.2.2. User system requirements

Table 6. System requirements of Educational material Aggregator

	Windows requirements	Apple requirements	Linux requirements
Operating system	Windows XP SP2+ Windows Vista Windows 7 Windows 8 or later	Mac OS X 10.6+ iOS 4+	Ubuntu 12.04+ Debian 7+ OpenSuSE 12.2+ Fedora Linux 17
Software	IE 11+, Chrome 17+, Firefox 16+, Safari 6+, Opera 15+		
Processor	Any x86, x64 or ARM v7 processor at 1Ghz and above		
Free disk space	80 MB		
RAM	512 MB		
Display resolution	From 320x240 to 1920x1080		

[38] http://getbootstrap.com/

[39] github.org/telemed-duth/carre-edu

A social account or standard register procedure is required for the rating function to be enabled. The search and viewer module are available without authentication.

6.2.3. Code metrics

The project is open source using the MIT License (MIT) and is hosted on github.org/telemed-duth/carre-edu.

The current release is v0.2 and the source code can be obtained from https://github.com/telemed-duth/carre-edu/releases/tag/0.2. The current implementation utilizes packages for the server backend component using the NPM package manager and the Bower package manager for Frontend component. All libraries are open source and current versions are available using the hyperlinks below.

Table 7. Used external libraries statistics

Package Manager	Libraries statistics
Server backend NPM	89 server libraries http://beta.carre-project.eu:9999/package.json
Frontend side Bower	20 client packages http://beta.carre-project.eu:9999/bower.json

Code quality analysis is based on lines of code and number of files as well as code complexity of each file which is previewed live at beta.carre-project.eu:9999 , as shown in Figure 23.

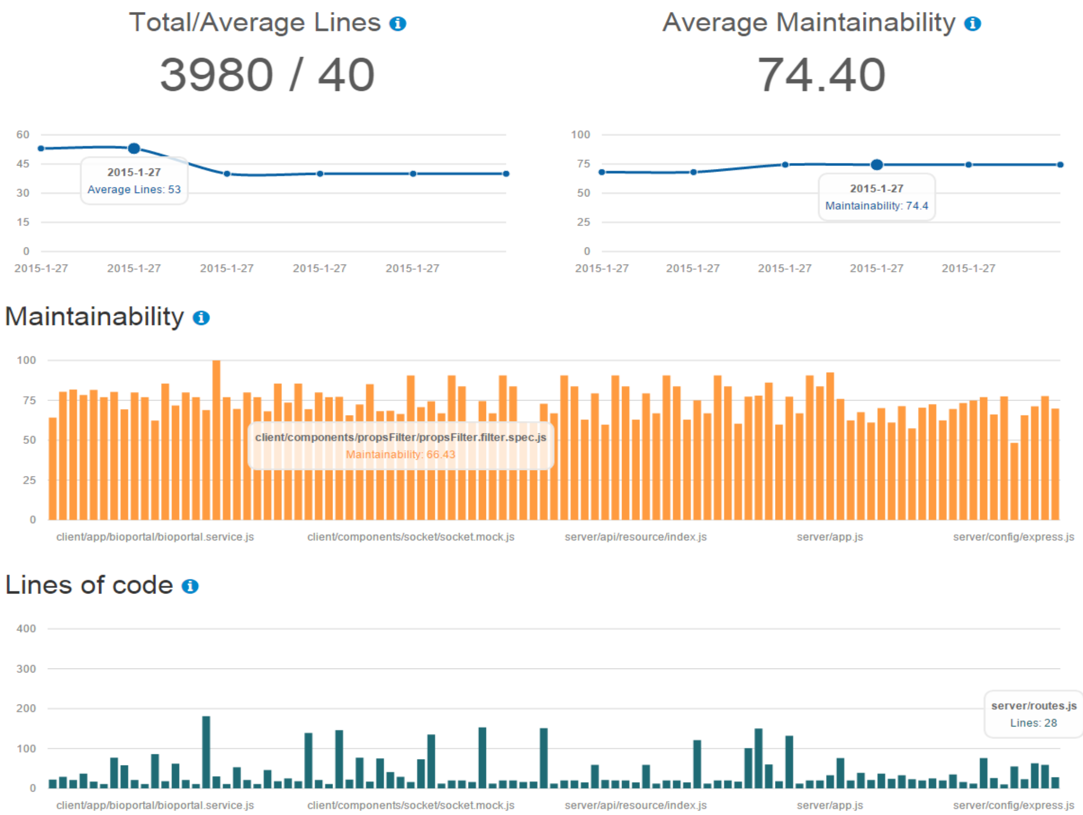

Figure 23. The code metrics are generated live using Plato[40].

[40] https://github.com/es-analysis/plato

7. Conclusion

This report summarised the data aggregation for medical evidence and educational data. The tasks listed in T3,4 in DOW have been achieved. The key functional blocks of the aggregator system detailed in Figure 2 and Figure 20 have been implemented, and the code can be found in the zipped file that submitted along with this report. As stated in the document, some of the tasks in Figure 2 will be done along with other tasks. In the integration stage in WP7, the services will be tested and improved, and a revised version of the software will be submitted.

Annex 1
Medical Evidence Data Aggregator Software

What is CARRE Medical Evidence Data Aggregator?

The main goal of the CARRE Medical Evidence Data Aggregator is to gather medical knowledge with the aims to 1) enrich the evidence of the existing risk descriptions as entered manually by medical experts and 2) identify new risk associations for cardiorenal diseases and comorbidity as published in medical literature during and beyond the project's lifetime. The evidences data, once confirmed by the medical experts, will be stored in the CARRE semantic RDF repository, which will be linked to open linked data repository to be used by public.

There are two parts of the aggregator which fulfills the above mentioned two goals, they are: the **Known Medical Evidence Data Aggregator** and the **New Risk Association Data Aggregator**.

– The **Known Medical Evidence Data Aggregator** is a Web service which allows expert to verify and evaluate the new evidence for the known risk associations based on the data mining results by the beackend system.
– The **New Risk Association Data Aggregator** is a backend service that identify the new risk associations.

Download

Known Medical Evidence Data Aggregator:

– Source (178 KB): CARRE_D.3.4_Aggregators_EvidenceData_Software_KnownRisk.7z (Java code)
download from http://www.carre-project.eu/innovation/medical-evidence-aggregator/

New Risk Association Data Aggregator

– Source (25 MB): CARRE_D.3.4_Aggregators_EvidenceData_Software_NewRisk.zip.zip (Java code)
download from http://www.carre-project.eu/innovation/medical-evidence-aggregator/

Medical Evidence Data Aggregator is Open Source

CARRE Medical Evidence Data Aggregator is Open Source and can be freely used in Open Source applications under the terms GNU General Public License (GPL).

Annex 2
Risk Model Semantic Data Entry System

What is CARRE Risk Model Semantic Data Entry System?

The Risk Model Semantic Data Entry system was initially developed in order to capture the risk associations identified in D.2.2. The Drupal content management system[41] has been customised to reflect the structure of the model presented here, so that observables, evidence sources, risk elements and associations can be entered via web forms, and automatically translated to RDF.

Visit

The Risk Model Semantic Data Entry System v1.0 is

available at http://carre.kmi.open.ac.uk

The CARRE Risk Model Semantic Data Entry System is Open Source

[41] https://www.drupal.org/

Annex 3
Educational Resource Aggregator Software

What is CARRE Educational Resource Aggregator?

The aim of the educational resource aggregator is to harvest educational resources from 3rd party repositories, present these to the medical expert for annotation and rating, and output the results of the annotation (together with resource metadata) to the CARRE public RDF repository.

The main parts of this aggregator are: the **Resource Retriever,** the **Resource Rating,** the **Resource Metadata Processing,** and the **User Application**.

- The **Resource Retriever** accepts CARRE concept terms from the CARRE public RDF repository and uses them to formulate queries to external 3rd party educational resource repositories. The results of this search are parsed to extract metadata. Then the retrieved results and metadata are displayed to the expert user for rating and annotation (via the aggregator front end). The module consists of 2 services that make use of SPARQL protocol in the case of query term extraction from the CARRE server and API requests to each educational repository.
- The **Resource Rating** module allows the input of expert user opinion and annotation, and also calculates subjective scores that measure the quality of the resource. Expert rating involves assessment of content-keyword relevance, content accuracy and depth of coverage, while the automatic systems rating is based on the Readability Test of the Flesch-Kincaid algorithm, and rating based on the latest modified version of the article and number of revisions. The module is an optional process that requires special user privileges. Each user role is assigned to different rating criteria according to user authorization. As of version 0.2, only 2 user roles have been taken into implementation: (a) the expert, and (b) the general public.
- The **Resource Metadata Processing** module involves metadata enrichment via semantic web sources (such as NCBO BioPortal medical ontologies and DBpedia). The module is a combination of 3 services that collect data per article, making multiple SPARQL requests to enrich the data and finally store it as a unique identified resource into the local MongoDB datastore. Then data is transformed into RDF triples in order to be inserted to CARRE educational repository.
- The **User Application** is a web application accessible at http://edu.carre-project.eu/. The visible components of the web application are built upon html5 and CSS3 using modern frameworks for consistency and responsiveness like Twitter Bootstrap[42] CSS Framework have been extensively used. Bootstrap is also responsible for mobile/tablet view of the web application.

Visit

Educational resource search and rate application:

visit at: http://edu.carre-project.eu

or at: http://www.carre-project.eu/innovation/educational-data-aggregator/

Download

Educational resource aggregator v0.2:

- Source (117 KB): CARRE_D.3.4_Aggregators_MedicalEvidence_Educational_v0.2.zip

  download from http://www.carre-project.eu/download/software/d.3.4_aggregators_medicalevidence_educational/CARRE_D.3.4_Aggregators_MedicalEvidence_Educational_v0.2.zip

  or from http://www.carre-project.eu/innovation/educational-data-aggregator/

[42] http://getbootstrap.com/

Deploy on your own server

Minimum Requirements: 1GB RAM + 1GB HDD The deployment is supported only on a unix* like machine (linux , Mac) and requires the following libraries to installed on your computer:

- NodeJS application sever
- MongoDB database server
- Git version control system

Next you should clone the repository at github, install all dependencies and run the build script.

```
$ git clone https://github.com/telemed-duth/carre-edu.git
$ npm install -g bower grunt-cli
$ npm install && bower install
$ grunt serve
```

Educational Resource Aggregator is Open Source

CARRE Educational Resource Aggregator is Open Source and can be freely used in Open Source applications under the terms MIT License (MIT).