

DIGIT.B4 – Big Data PoC

RTD – Health papers

D02.01 PoC Requirements

everis Spain S.L.U

Table of contents

1 Introduction	5
1.1 Context	5
1.2 Objective	5
2 Data Sources	6
2.1 Selecting the data	6
2.2 Obtaining the data	6
2.3 Filters and fields	7
2.4 Modelling taxonomies	8
2.5 Creating Dictionaries	10
3 Data presentation and visualization	11
3.1 Sections	11
3.1.1 Home and About	11
3.1.2 Data	12
3.1.3 Analysis	14
3.2 Possible evolutions	18
4 PoC Requirements Summary	19

Table of Tables

Table 1- Selected sources to use in the PoC	6
Table 2 - Estimated number of registries in PubMed.....	6
Table 3 - Estimated number of registries in CORDIS	7
Table 4 - Filters used to select the registries to analyse in the PoC	7
Table 5 - Fields selected to use in the PoC.....	8
Table 6 - Health category dictionary.....	9
Table 7 - Research activity dictionary	10
Table 8 - Research activity dictionary	10
Table 9 - Requirements summary	20

Table of Figures

Figure 1 - Web page structure	11
Figure 2 - Web page view	11
Figure 3 - Home section.....	12
Figure 4 - About section.....	12
Figure 5 - Data section.....	13
Figure 6 - Advanced search section.....	¡Error! Marcador no definido.
Figure 7 - Advanced search example.....	¡Error! Marcador no definido.
Figure 8 - Temporal discovery	15
Figure 9 - Temporal discovery example	15
Figure 10 - Research analysis (bubbles).....	16
Figure 11 - Research analysis (lines).....	16
Figure 12 - Term analysis (Wordcloud)	17
Figure 13 - Term analysis (bars)	17
Figure 14 - FCA section	17

1 INTRODUCTION

1.1 Context

The execution of this proof of concept, in cooperation with DG RTD, will showcase the use of big data in the EC research domain and prove the usefulness and policy benefit that big data can bring. This proof of concept shall demonstrate the use of text mining techniques used on large amounts of unstructured research papers as a means to identify areas of interest in research, to be considered as additional input prior to launching calls for grants.

In order to define the scope of the PoC, we have reviewed the user needs and assumptions and we have elaborated this document, which includes the following sections:

- Data Sources.
- Data presentation and visualization.

1.2 Objective

The aim of this document is to describe all the requirements that the PoC should fulfil in order to ensure the achievement of the objective of the project.

2 DATA SOURCES

The data is needed for the creation of a model suitable for the classification of documents within a domain, in this case biomedical and health literature.

2.1 Selecting the data

After analysing multiple data sources (Science Direct, PLOS one, the European Patent Office, etc.), the following ones have been selected due to the advantages shown below:



Data source	Advantages
 PubMed	Multiple filters (date, author, etc.); CSV, XML and TXT search extraction; massive amount of biomedical literature. This data source fills the need for real research information.
 CORDIS	Multiple filters (date, author, etc.); CSV, XML and TXT search extraction. This data source fills the need for real European funding information.

Table 1- Selected sources to use in the PoC

2.2 Obtaining the data

One of the most important steps in the PoC is the acquisition of data. In order to optimize resources and facilitate the data loading process, the CSV format has been selected due to its lightweight and simple structure. The PoC uses text to analyse and categorise the data. This text has to be long enough for the model to deduce its meaning, but not so long that the model has too much information and has to decide between lots of interpretations. The abstract is the perfect match, so the obtained data has to contain this field regardless of its source.

- **PubMed:** the data will be downloaded directly from the PubMed website (<http://www.ncbi.nlm.nih.gov/pubmed>) as XML. The ideal option would have been CSV, but CSV exports from PubMed do not contain any information about the abstract of the papers, which is a very important field for the classification model (the abstract is the text that summarises a project content, and the primary source of the model to perform the classification).

In order to facilitate the classification task, a converter between the PubMed XML export format and a CSV equivalent file will be implemented.

Without applying any other filters than “language” and “date created”, these are the estimated amount of records obtained:

Language	Date	Number of registries
English	2015/08/01 to present	≈ 100.000
English	2014/01/01 to present	≈ 2.000.000
English	2000/01/01 to present	≈ 11.000.000

Table 2 - Estimated number of registries in PubMed

- **CORDIS:** the CORDIS web export is limited to 5000 results, and the export files have the same drawbacks than those from PubMed: in CSV format the field “abstract” does not exist. However, all the projects from the FP6 and FP7 programmes, as well as some of the more recent H2020 projects can be downloaded as CSV from the European Union Open Data Portal (<https://open->

data.europa.eu). These files have the “subjects” field, which is equivalent to the “abstract” one.

An estimation of the amount of project records obtained selecting only the ones from the “Health” topic, written in English and from 2002 to present, can be made with the search results of the CORDIS webpage (<http://cordis.europa.eu/>):

Language	Programme	Date	Number of records
English	FP6-LIFESCIHEALTH, FP7-HEALTH, H2020-HEALTH	2002 to present	≈ 1800

Table 3 - Estimated number of registries in CORDIS

2.3 Filters and fields

The raw data has to be adjusted to the needs of the PoC. This can be achieved by applying some filters to narrow the classification process and selecting the most useful fields for the visualizations.

▪ Filters

In order to limit the number of registries in PubMed and avoid the “noise” (registries that fall out of topic but inside generic filters and distort the result of the classification model), the following filters have been defined, narrowing the number of registries to a more reasonable one:

Language	Date Created	Text availability	MeSH Terms	Number of records
English	2009/01/01 until present	abstract	Humans; In Vitro Techniques; Animal Use Alternatives; Animal Experimentation; Disease Models, Animal; Animal Testing Alternatives	≈ 3.000.000

Table 4 - Filters used to select the registries to analyse in the PoC

MeSH Terms specific filters have been defined to include basic scientific research that may be omitted by the generic “Species” filter set initially to “human”. This filter is very restrictive, and research topics that are not human-specific are omitted, even though they may be important to human research.

The field “Date Created” narrows the data to projects created between 2009/01/01 and present. These dates have been chosen for two reasons:

1. To match the PubMed project dates to the H2020 (Horizon 2020) programme. This will facilitate the analysis to compare the investigation trends with the CORDIS funding ones. The years before 2014 have been added to train the classification model (the training process is the action of fetching a model with a given amount of processed information so it learns how to classify new data on its own).
2. The objective of the PoC is to evaluate trends so the European research programmes target the right topics. These trends change very quickly over time, so if relatively old projects are analysed, the result may not fit the present needs of the field.

The resulting amount of records applying the previous filters, about 3.000.000, is more than enough to perform the analysis based on the experience that the team has with similar projects. This experience tells us that 200.000 records are enough to perform the modelling.

▪ Fields

To integrate the PubMed and the CORDIS registries, common fields in both datasets must be defined. These fields will be determined according to the value that each one can provide to the classification process and the visualization pages:

Field name	Description	In PubMed	In CORDIS
date_created	The creation date of a document in DD/MM/YYYY format	✓	✓
authors	A comma separated string with all the names of the document authors	✓	✓
title	The title of the document	✓	✓
abstract	The abstract text of the document	✓	✓
journal_iso_code	The ISO code belonging to the journal in which the article is published	✓	✗
keywords	The keywords associated with the article	✓	✗
url	The URL of the article	✓	✓

Table 5 - Fields selected to use in the PoC

Even though the “journal_iso_code” and the “keywords” fields are not present in the CORDIS data, they will be included for the PubMed registries because of their usefulness in the visualization sections.

2.4 Modelling taxonomies

The classification model needs a way to associate its results with real categories, so different taxonomies must be generated to achieve this goal.

The scope of this PoC focuses on health related topics, and the “Health Research Classification System” or “HRCS” (<http://www.hrcsonline.net/>) will be used to match the model outputs with health categories and research activities. This classification system has been widely adopted by research funders and offers a perfect starting point for the taxonomies modelling.

Other classification systems have been studied (for example the MeSH vocabulary, used by the PubMed article database), but even though they offer a perfectly valid way to classify data, adding multiple classification systems may be counterproductive due to the lack of value they add to the main objective of the PoC. However, this can be taken into account for a future evolution of the PoC.

To match the model results with a valid category and/or research activity, its contents must be adapted and expanded to transform the HRCS framework to an input that the model can understand. The adaptation focuses on creating structured datasets with the information defined in the framework. A set of keywords will be added to the framework for each category/research activity. This will help the model to accurately classify the documents analysed. These keywords will be validated by several field-specific experts.

▪ Health Categories Taxonomy

The HRCS framework has 21 separate categories that encompass all diseases, conditions and areas of health¹. Each category provides a description of the kind of affections enclosed in it. This description is what the experts will use to define the keywords² that will help the model to classify its results.

This taxonomy will be a 2-column CSV dictionary, as shown in the following example:

category_name	keyword
blood	blood
blood	clot
blood	platelet
...	...
cancer	leukaemia
cancer	cancer
...	...

Table 6 - Health category dictionary

▪ Research Activity Taxonomy

The HRCS framework has 8 basic research activity codes that are divided into 48 narrower categories.

For the model to be as precise as possible, each subcategory should be associated with a set of very specific keywords. The problem with the HRCS Research Activity codes is that even though its categories are all conceptually different, some of them refer to the exact same field of knowledge and their keywords are almost the same. For example, the categories “Marker Discovery” and “Marker Evaluation” refer to different concepts: the first one is about “discovery” and the last one about “evaluation”; but their field of knowledge (“markers”) is the same, thus their keywords are too. If a text containing marker’s keywords that would fit into the “Marker Discovery” category is classified this way, both categories (“Marker Discovery” and “Marker Evaluation”) will be assigned to the text, even though the “evaluation” one would not fit the meaning of the paper.

In order to avoid this problem, categories and subcategories whose keywords are the same may be merged into broader ones.

The classification will take place in two stages:

¹ <http://www.hrcsonline.net/hc>

² The keywords in this document are illustrative, not real examples.

1. The clusters will be classified with the 8 (or less, depending on the merging process) broader categories

broad_research_activity	keyword
underpinning	underpin
underpinning	build
underpinning	construct
...	...
prevention	prevent
prevention	avoid
prevention	inhibit
...	...

Table 7 - Research activity dictionary

2. Each classified cluster will be classified again with the narrower categories of the corresponding parent category. For example, the dictionary for the underpinning category should have the following structure:

narrow_research_activity	keyword
biological	gene
biological	molecule
biological	cell
...	...
psychological	perception
psychological	cognition
psychological	learn
...	...

Table 8 - Research activity dictionary

▪ OECD Taxonomy

The OECD taxonomy is based on an internal document with keywords for different scientific fields. Its structure is exactly the same as the one defined in “Health Category Taxonomy”.

2.5 Creating Dictionaries

To clean the text of undesired terms and provide a complete list of the health terms to look for when the classification is performed, the creation of several dictionaries is needed.

A dictionary is a structured n-column file where data such as stopwords, synonyms, hyperonimics, terms and several of its properties are listed. A more extensive description of these structures is available in the section 3.3 of the “D02.02.Technological_Architecture_v1.1.docx”

The terms in these dictionaries have been validated by everis and Fundación Ramón Domínguez (<http://www.fundacionramondominguez.es>) experts.

3 DATA PRESENTATION AND VISUALIZATION

The results of the analysis will be published in a web page with a standard content distribution:

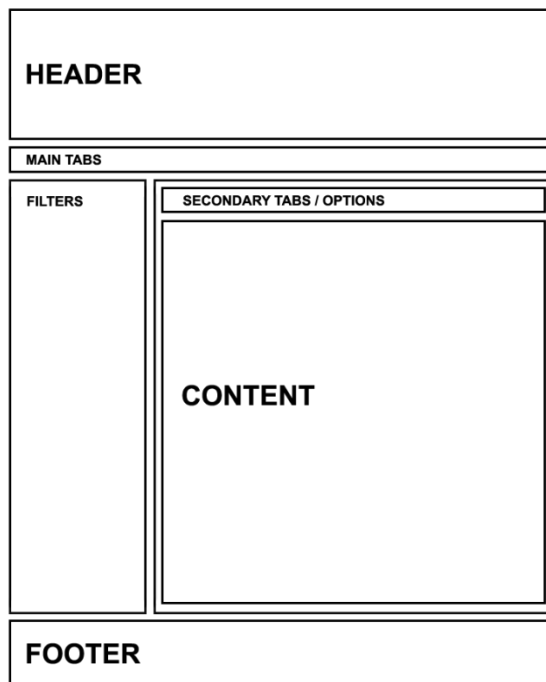


Figure 1 - Web page structure



Figure 2 - Web page view

Four main sections will be designed: **HOME**, **DATA**, **ANALYSIS** and **ABOUT**; all of them accessible through the MAIN TABS panel of the interface.

3.1 Sections

3.1.1 Home and About

Both the **HOME** and **ABOUT** sections are static web pages. The first one simply provides a more visual way to access the web contents, whilst the second one explains the analysis process and the objectives of the PoC to the users.

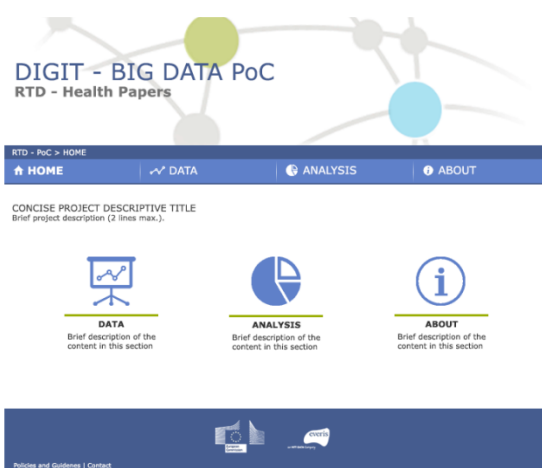


Figure 3 - Home section



Figure 4 - About section

3.1.2 Data

In the DATA section, the user will be able to list all the papers involved in the analysis. This section is composed of two pages: the “DATA” page, where the content of the analysis is listed, and the “Advanced Search” page, where the user can filter the results by different parameters.

In the “DATA” page, in addition to list all the papers used in the analysis, the user can filter this data by two basic parameters: “Health” categories and “Research Activity” categories and sub categories (at the left side panel).

The figures below show two examples one with all the categories related to Health (Figure 5) and other with all the categories related to “Research Activity” and the visualization of the subcategories related to “aetiology” and “disease management”.



Figure 5 – Health categories



Figure 6 - “Research Activity” categories and sub-categories

For example, if a user selects the “Metabolic” and “Neurological” categories from the health categories options at the left side panel (the bold grey ones in the Figure 5), only the papers classified this way will appear in the results. That is, only the papers which contain any keyword related to any of the categories selected (here “Metabolic” and “Neurological”) in the abstract will appear as part of the results lists



Figure 7 - Data section

Both these parameters are based on the Health Research Classification System (HRCS). The user will also be able to download a CSV or XSL file with the information displayed in this page from the link placed at the right side of the “SECONDARY TABS/OPTIONS” panel.

In the “Advanced Search” page, the user can narrow the result list by adding search filters such as “Current date”, “Journal”, “PoC category” (the categories generated during the analysis by the Formal Concept Analysis algorithm with the overlapped terms), “Title”, “Author” or “Terms”.



Figure 8 - Advanced search section

For example, in the Figure 9, only data created on 2011/09/03, published in the journal with ISO code “J Nat Sci”, belonging to the HRCS health category “Blood” and the PoC category “Blood cancer”, and whose abstract contain the term leukaemia would appear in the results.

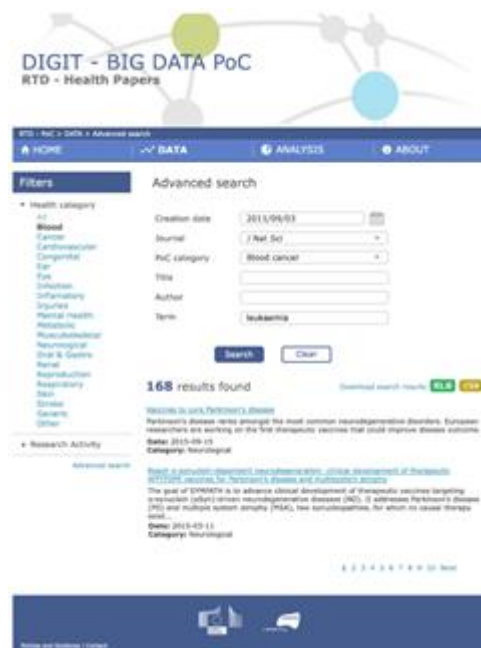


Figure 9 - Advanced search example

3.1.3 Analysis

The ANALYSIS section is where the functionalities fulfilling the main objectives of the PoC are implemented. It comprises four subsections: “Temporal discovery”, “Research Analysis”, “Term Analysis” and “FCA” (Formal Concept Analysis).

Each of these subsections contains a different way to interpret the analysis performed over the documents. The visualisation is designed to be as user friendly as possible using multiple types of charts and web applications, thus making the results clear and valuable for non-technical users.

In the “**Temporal discovery**” subsection, the tool will display a clusterization with the amount of papers belonging to each category selected in the left panel. The user will be able to select a date, making the chart show the number of papers in each category until that date. Additionally, by passing the mouse pointer above its corresponding cluster, the concrete number of papers per category will be displayed.

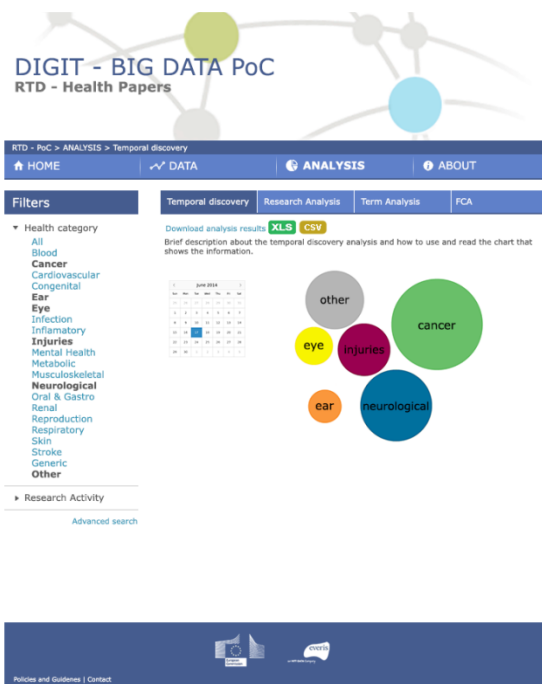


Figure 10 - Temporal discovery

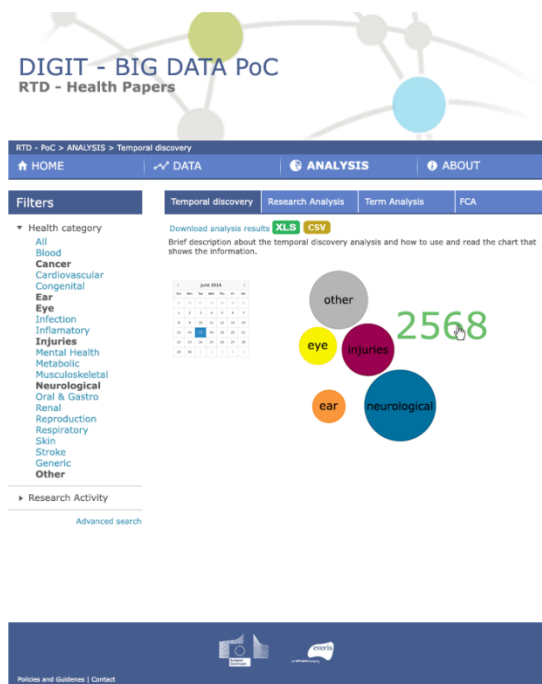


Figure 11 - Temporal discovery example

For example, in figure 10 the number of papers belonging to each category selected on the left panel (“cancer”, “neurological”, “other”, “injuries”, “eye” and “ear”) until the selected date (“2014/06/17”) is represented by the size of each bubble. In Figure 11, the number of papers falling under the “Cancer” category is showed.

The charts in the “**Research Analysis**” section show the count over time of articles in each HRCS category. There are two different charts:

- In the “bubble” chart, when the user holds the mouse pointer on top of one of the category circles (in this case the chart shows all the categories selected in the left side panel), the circles change to numbers and let the user see the actual figures of the analysis (as shown in Figure 12).
- The line chart³ on the other hand makes the comparison process easier to see, showing a dotted-line representation of the selected categories (Figure 13)

³ This chart may be confusing when lots of categories are selected due to the multiple lines drawn. The gap between the number of papers published by different categories can make the interpretation of results more complicated.



Figure 12 - Research analysis (bubbles)



Figure 13 - Research analysis (lines)

The “**Term Analysis**” tab contains charts that represent how many times a term appears throughout all the documents, filtered by the categories on the left panel. If the wordcloud is selected (Figure 14), by using the right side subpanel, the users can modify the amount of terms that appear and the minimum occurrences that a term should have to appear in the chart. Users can also search for a specific term to see how many times it appears in the analysis. The bar plot (Figure 15) shows a ranking with the top 10 terms.

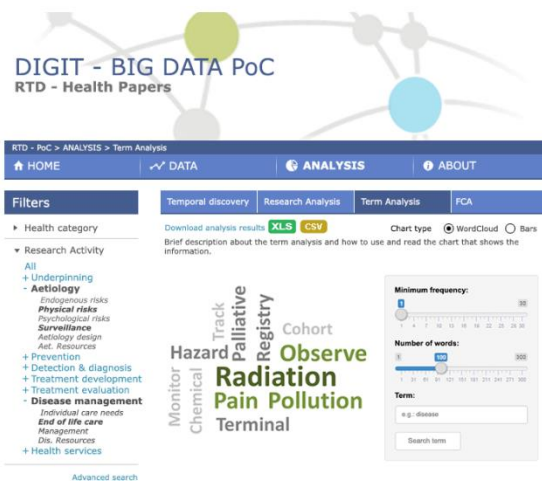


Figure 14 - Term analysis (Wordcloud)

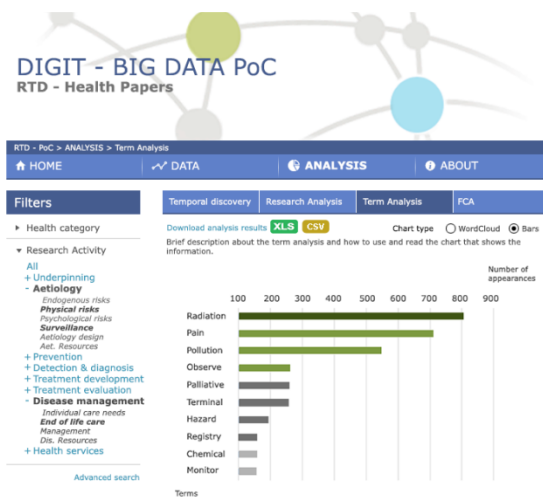


Figure 15 - Term analysis (bars)

The last ANALYSIS subsection, “FCA”, shows the relationship between categories and overlapping between them. These overlappings represent new categories and help to narrow the classification of medical papers based on a broader classification model.

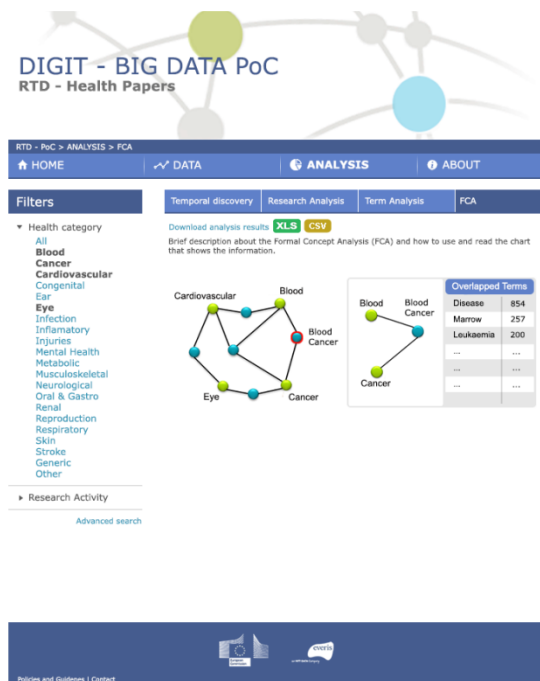


Figure 16 - FCA section

For example, in Figure 14, the FCA “Blood Cancer” category has been created with the overlapped terms showed on the right panel. These overlapped terms exist in more than one category, in this case “Blood” and “Cancer” becoming the foundation of the

new “Blood Cancer” category. The user will be able to select the overlapped dots and investigate the terms that it contains.

The users can download each analysis result in CSV and XLS format.

3.2 Possible evolutions

There are multiple data that can be taken into account when visualizations are defined.

One of the most interesting parameters that may provide very useful information is the amount of funding for the projects fetched from CORDIS. Even though for this PoC this information is irrelevant, if properly analysed it could be used to analyse how research budget is distributed and its relation with the trends identified by the PoC.

Also, if provided, impact factor and quartile information from the Web of Science Journal Citation Report can be added as extra filter parameters for the PubMed data.

4 POC REQUIREMENTS SUMMARY

The following list of requirements summarises the main actions a user will be able to perform in the PoC UI (user interface).

	Requirement	Description	Type
UR-01	List Papers	List all the papers used in the analysis	User Requirement
UR-02	Filter List of Papers	Filter the results displayed by "Health Category", "Research Activity", "Creation date", "Journal", "PoC category", "Title", "Author" and "Term"	User Requirement
UR-03	Download List of Papers as CSV	Download the resulting papers list in a CSV format file	User Requirement
UR-04	Download List of Papers as XLS	Download the resulting papers list in a XLS format file	User Requirement
UR-05	Visualize temporal evolution	Visualize the amount of papers of the list in each category until a selected date	User Requirement
UR-06	Download temporal evolution as CSV	Download the temporal evolution analysis in CSV format	User Requirement
UR-07	Download temporal evolution as XLS	Download the temporal evolution analysis in XLS format	User Requirement
UR-08	Visualize research analysis	Visualize the count over time of articles in each HRCS category.	User Requirement
UR-09	Download research analysis as CSV	Download the research analysis in CSV format	User Requirement
UR-10	Download research analysis as XLS	Download the research analysis in XLS format	User Requirement
UR-11	Visualize term frequency	Visualize the number of times a term appears in all the papers	User Requirement
UR-12	Download term frequency as CSV	Download the term frequency analysis as a CSV file	User Requirement
UR-13	Download term frequency as XLS	Download the term frequency analysis as a XLS file	User Requirement
UR-14	Visualize the overlapped categories	Visualize the new categories generated by the model based on the overlapping terms	User Requirement
UR-15	Download the FCA results as CSV	Download the FCA results as a CSV file	User Requirement
UR-16	Download the FCA results as XLS	Download the FCA results as a XLS file	User Requirement
FR-01	Obtain CORDIS data	Obtain the CORDIS projects data from 2013 to present	Functional Requirement
FR-02	Obtain PubMed data	Obtain the PubMed data from 2009 to present	Functional Requirement
FR-03	Create Health Categories Taxonomy	Based on the HRCS Health Categories, a taxonomy of its structure needs to be built for the classification model.	Functional Requirement

	Requirement	Description	Type
FR-04	Create Research Activity Taxonomy	Based on the HRCS Research Activity codes, a taxonomy of its structure needs to be built for the classification model.	Functional Requirement
FR-05	Create OECD Taxonomy	Based on internal knowledge, a taxonomy of the OECD classification structure needs to be built for the classification model.	Functional Requirement
FR-06	Create classification dictionaries	With the help of several health experts, dictionaries with technical terms will be built for the classification model.	Functional Requirement

Table 9 - Requirements summary