DIGIT.B4 – Big Data PoC

RTD – Health papers

D01.04.Final Project Report

everis Spain S.L.U

Table of contents

Table of figures

1 PROJECT INFORMATION

The summary, objectives, justification, inventory of deliverables and baseline are described in the following points.

1.1 Project Summary

This project showcase the use of big data in the EC research domain and prove the usefulness and policy benefit that big data can bring

This project also demonstrates the use of text mining techniques used on large amounts of unstructured research papers as a mean to identify areas of interest in research, to be considered as additional input prior to launching calls for grants.

The information to analyse is gathered from two data sources of biomedical and health literature. These data sources are PubMed and Cordis. The information is shown in the tool by various graphs.

1.2 Project Objectives

This project is part of the ISA Action1.22 – Big Data and Open Knowledge for public administrations. The ISA action's objective are the following:

1. To identify the requirements and challenges public administrations in Europe are confronted with in the area of big data and open knowledge and identify opportunities.
2. To identify best practices by public administrations and/or organisations which could be used as lessons learnt including an assessment of the tools and solutions that these best practices have implemented.
3. To identify synergies and areas of cooperation with the policy DGs and the MSs in the big data and open knowledge domain.
4. To identify areas of interests whereby the ISA programme and its proposed successor could have an active role in launching initiatives for enabling practical concrete implementations that will answer the requirements of the public administrations in Europe.

This project contributes to the point 3 above, as its objective is to execute a proof of concept, in cooperation with DG RTD, which proves how big data techniques can be applied in the research domain and to demonstrate the policy benefits 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 of identifying areas of interest overlap that a particular research area should consider prior to launching calls for grants.

1.3 Project Justification

The project justification is to manage a project with two main activities:

1. To provide consultancy on Big Data Technologies and support DIGIT on the delivery of formative and consultancy actions.
2. To lead a PoC to demonstrate the use of Big Data in the context of text analysis for DG RTD:
    a. To capture and formalise the requirements for the PoC, as expressed by DG RTD.
    b. To develop and host an information system implementing the requirements.

1.4 Inventory of Project Deliverables

This section includes the inventory of the identified project deliverables that have been developed.

Deliverables		
Deliverable Code	**Name of Deliverable**	**Date of Acceptance**
D02.01	Poc_Requirements	29/12/2015
D02.02	Technological_Architecture	08/10/2015
OD02.01	Big_Data_Technological_Reference_Model	07/10/2015
D03.01	Data_Linguistic_Understanding	10/12/2015
D03.02	Dictionaries	29/02/2016
D03.03	Text-mining_Models	06/04/2016
D04.01	Information_System	29/02/2016
D04.02	User_manual	19/02/2016
D04.03	Migration_plan	29/02/2016

1.5 Project Baseline

The tasks were grouped into four main packages to make a planning of the project and were distributed in sub-tasks. The four blocks of tasks are:

- Task 01 – Project Management
- Task 02 – Requirements Analysis
- Task 03 – Text Mining
- Task 04 – Publication of results

The figure below shows the initial division of the tasks:

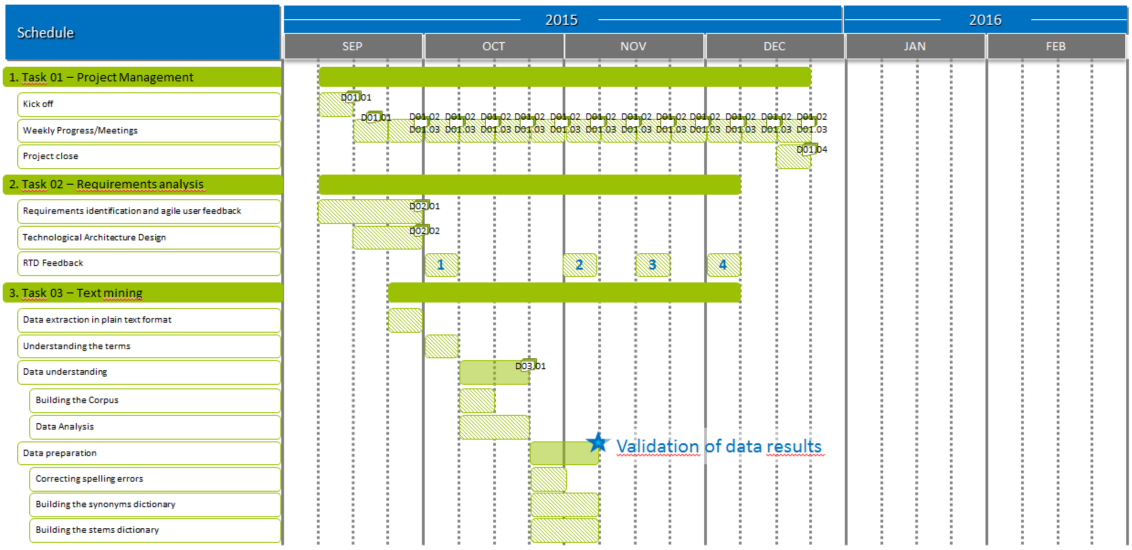

Figure 1 - Initial planification Task 01, Task 02 and Task 03

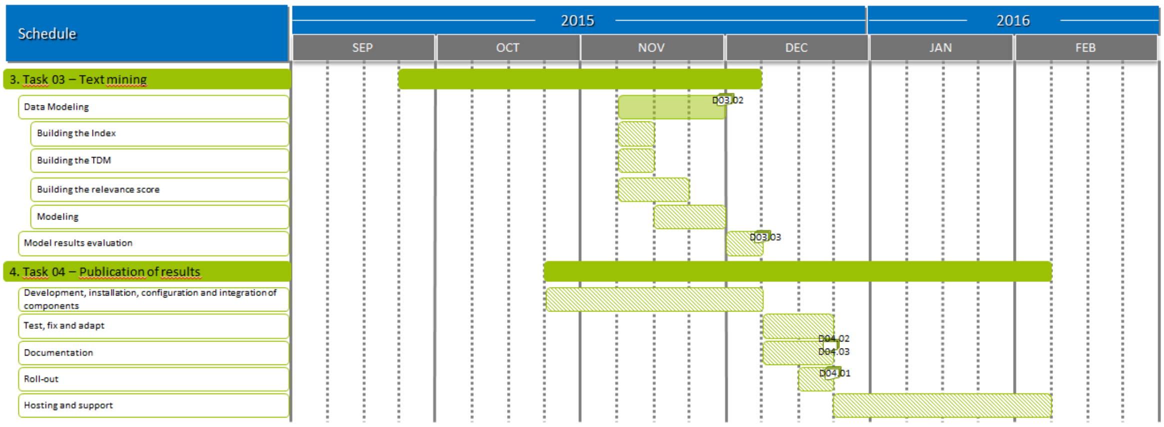

Figure 2 - Initial planification Task 03 and Task 04

The project was initially planned to have duration of five months, starting the fourth week of September 2015 and ending the first week of February 2016. Finally due to some issues, the project has had duration of five months and two weeks ending the third week of February. The main delay has occurred in the tasks related to task 03 that and task 04 where each one was lengthened by one week.

The final plan can be seen in the following figures:

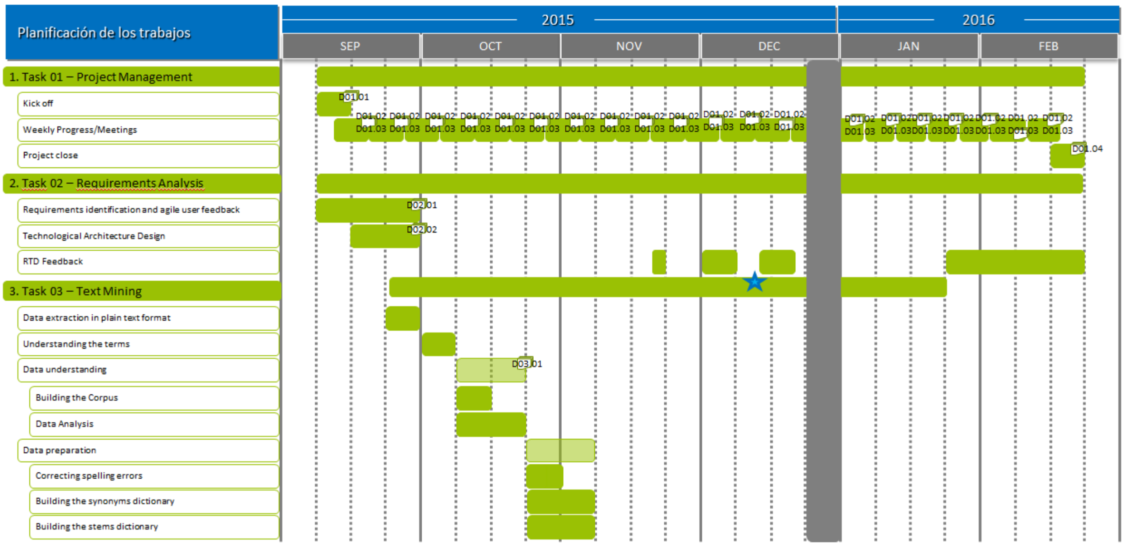

Figure 3 - Final planification Task 01, Task 02 and Task 03

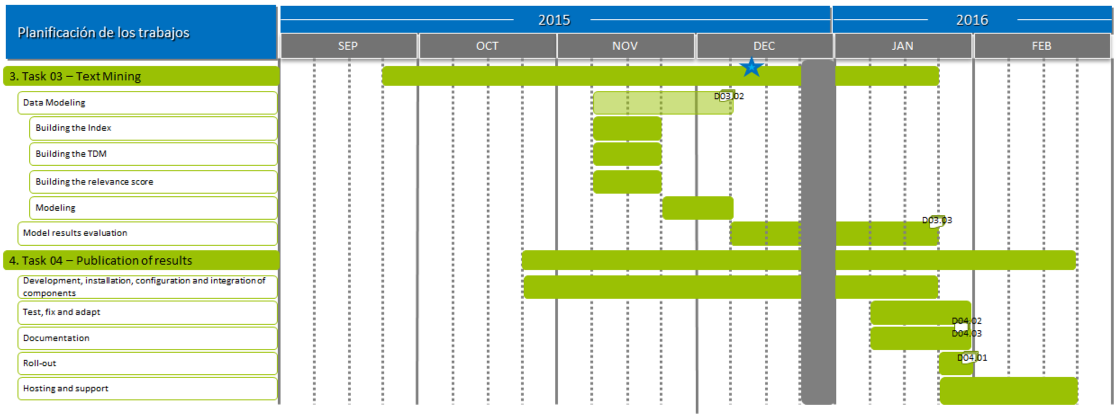

Figure 4 - Final planification Task 03 and Task 04

2 ISSUES AND RISKS

This section shows the problems identified during the project implementation and the solutions applied to solve them.

Issues/Risks		
Issue/Risk ID	**Description**	**Solution**
1	Not having access to CORDIS database on time	PO allows everis to access to CORDIS database.
2	everis tells DIGIT.B4 that a part of the CORDIS project (presentation and visualization) is in charge of another everis team	everis asks CORDIS everis team directly for the needs
3	The late feedback from RTD related to some requirements will cause delays	DIGIT.B4 send an email to DIGIT RTD asking for feedback about the requirements
4	Alignment with the visualization requirements expected by RTD	everis deliver the D02.01 whit one week delayed to insert figures with the concrete and real data from RTD (like examples)
5	The late feedback from RTD related to the categorization will cause stops in the execution of some tasks	DIGIT.B4 send an email to DIGIT RTD asking for feedback about the categorization
6	everis says that the generation of none expecting results by RTD will cause delays	DIGIT.B4 sends an email to DIGIT RTD asking for feedback about the generation of expecting results.
7	Include the new visualizations related to temporal evolution of keywords will cause delays	everis includes the new visualizations with one week delayed. DIGIT.B4 and everis agree to show the tool by implementing a demo to RTD in order to verify the correct development of it.
8	Not understand of the real scope of the project by RTD will cause not the expected results and will cause delays	everis explains better to RTD the scope of the project. DIGIT.B4 and everis agree that the final results will be available in late January.
9	Include the new modifications related to improve the tool (date range, select/deselect all and horizontal keywords cloud) will cause delays	everis includes the new visualizations with two weeks delayed

3 NEXT STEPS

This section shows the next steps identified for the following waves of the tool:

- Define the new requirements for the tool such as:
  - Modify the Data Section to include the possibility of searching by a period instead a concrete date
  - Insert the possibility of getting all the information at the Analysis section but by filtering out per ad-hoc terms
  - Translate the filters insert at the Data Section into the analysis section
  - Insert the possibility of filtering by source
  - Insert the possibility of the automated feeding of data
  - Insert the possibility at the Temporal evolution per term to filter by concrete terms