DIGIT.B4 – Big Data PoC

GROW – Transpositions

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

The application of this proof of concept (PoC) started with the purpose of showing the benefits and feasibility of the application of techniques of text mining to support compliance checks conducted by Commission staff on national transposition measures EU directives .

To do this, in the context of the PoC, the checks were performed on the basis of national texts communicated to the Commission by the Member States of their national measures transposing the Directive 2011/7/EU. This Directive was analysed in four languages: English, French, German and Spanish.

After analysing the results, it was shown that the techniques of text mining are not optimal for the use case of this proof of concept , so other techniques to achieve the results were explored, in this case development an algorithm that provided more optimal results than the techniques of text mining.

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 objectives 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 2 above, as its objective is to execute a proof of concept, in cooperation with DG GROW, whose aim is to prove how big data techniques can be applied to support the compliance checks carried out by Commission staff on national measures transposing EU directives.

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 GROW:

    a. To capture and formalise the requirements for the PoC, as expressed by DG GROW.

    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	12/05/2016
D02.02	Technological Architecture	07/04/2016
D03.01	Data Linguistic Understanding	01/07/2016
D03.02	Dictionary	01/07/2016
D03.03	Text mining Models	01/07/2016
D04.01	Information System	01/07/2016
D04.02	User Manual	01/07/2016

1.5 Project Baseline

The tasks were grouped into four main blocks when making the project plan, which were divided in to several subtasks. 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:

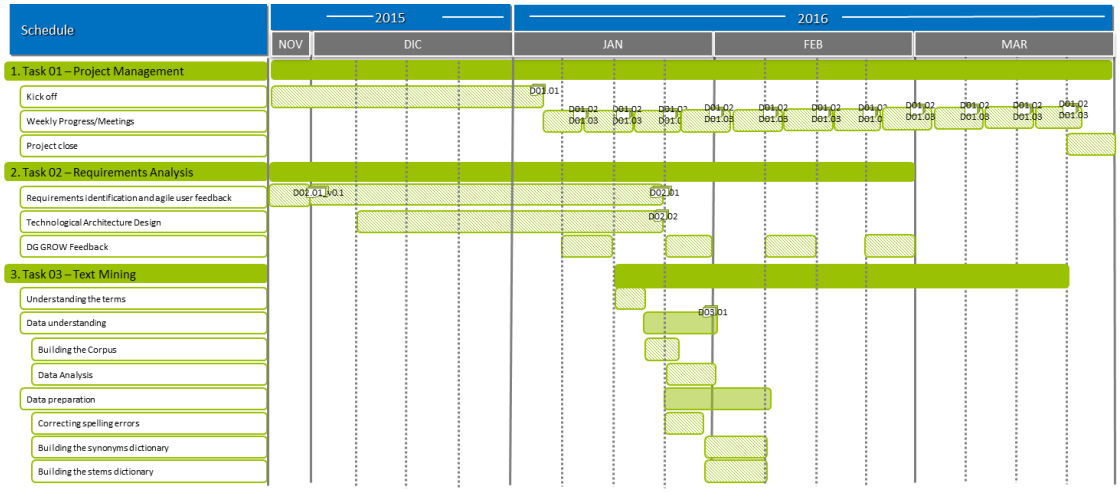

Figure 1 - Initial planning Task 01, Task 02 and Task 03

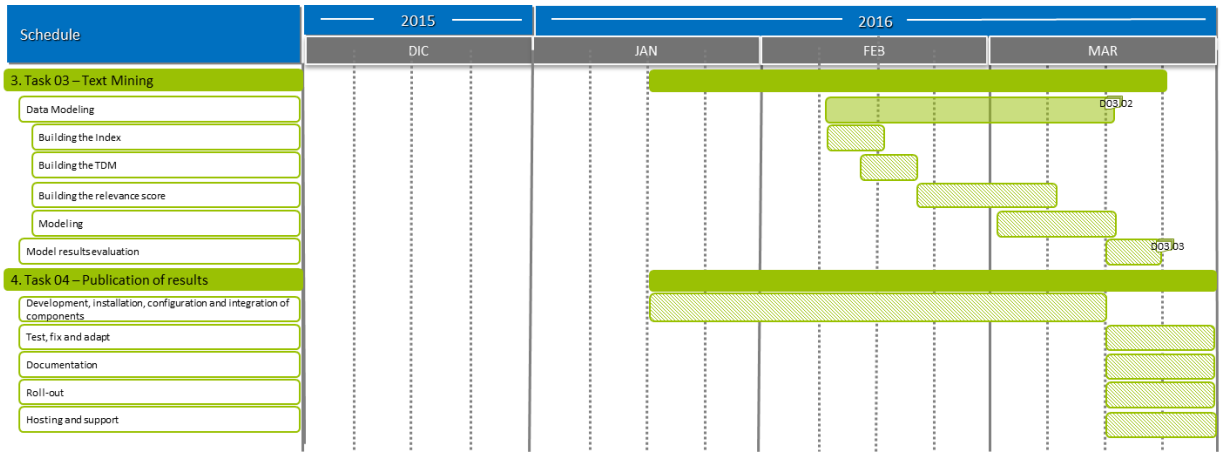

Figure 2 - Initial planning Task 03 and Task 04

Initially, the planned duration of the project was of four months and one week, starting the fourth week of November 2015 and ending the fourth week of March 2016. In the end, the project has had duration of six months and one week, ending the fourth week of May. Several improvements were made in the algorithm between April and May, in order to optimize the project's results. This meant that the date of completion of the project was later than initially expected.

The final plan can be seen in the following figures:

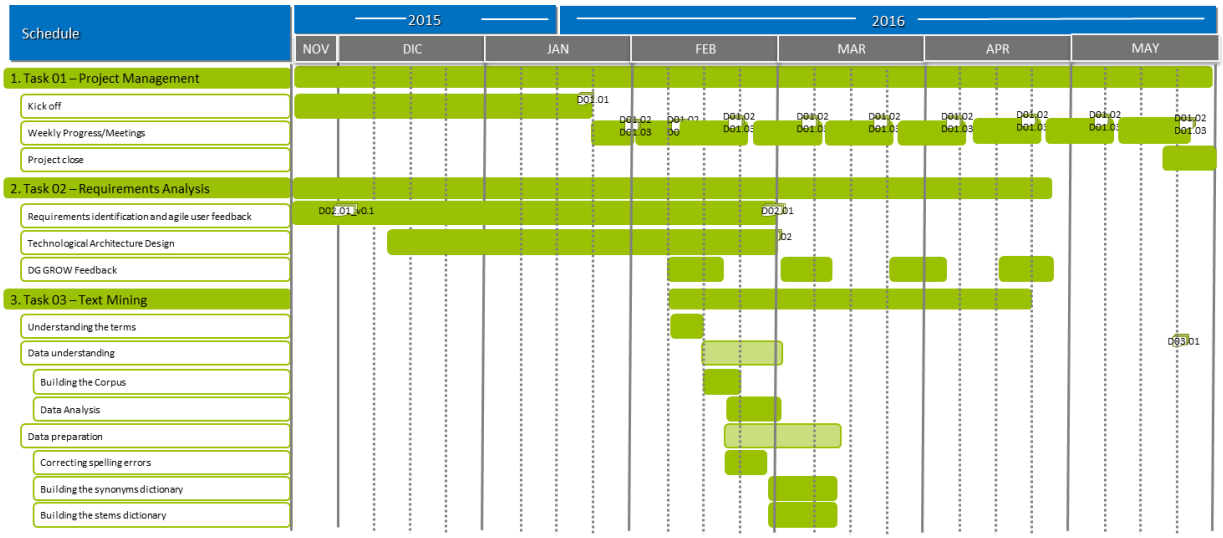

Figure 3 - Final planning Task 01, Task 02 and Task 03

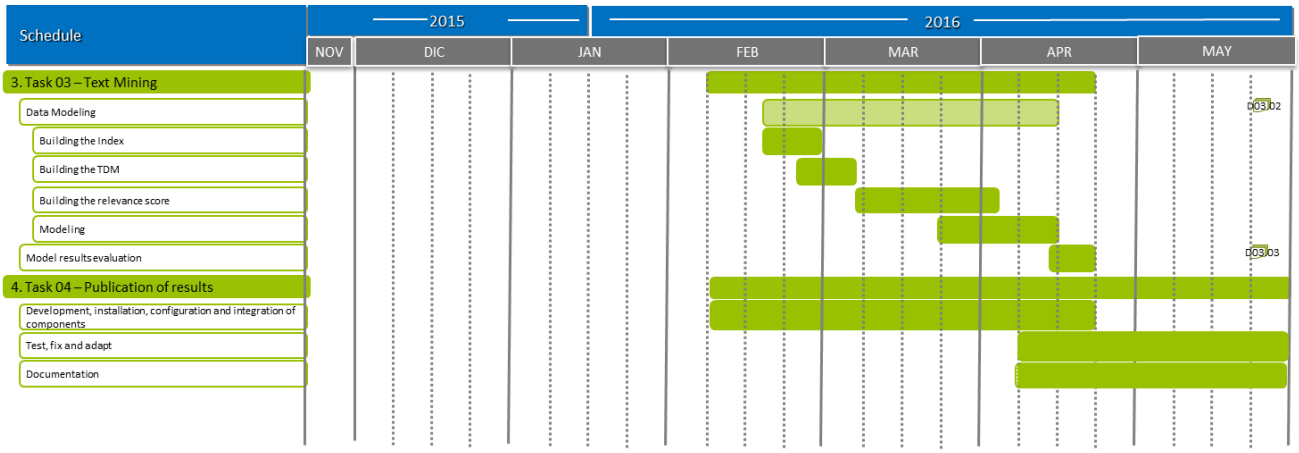

Figure 4 - Final planning Task 03 and Task 04

2 ISSUES AND RISKS

This section shows the problems identified during the project implementation and the solutions employed to overcome them.

Issues/Risks		
Issue/Risk ID	**Description**	**Solution**
1	The late feedback from DG GROW related to the validation of the first draft version of D2.01 could cause delays	To establish suitable deadlines for the documents' revisions, to avoid compromising the project's development.
2	The not alignment with the requirements expected by GROW	To organise periodical meetings in order to show the project's progress, to mitigate these risks.
3	The late feedback from DG GROW related to the validation of D2.01 could cause delays	To organise periodical meetings in order to show the project's progress, to mitigate these risks.
4	Not receiving similar texts of late payment on time	To identify the external parties involved in the project and to establish procedures beforehand to collect all the necessary information
5	The information getting from the text-mining techniques is not as good as expected	To develop the project by phases, establishing checkpoints in order to assess the project's evolution and the suitability of the results obtained.

3 NEXT STEPS

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

Due to the complexity of the objectives of the PoC and the non-ideal scenario for applying text-mining techniques, as highlighted in the conclusions of the data linguistic understanding document (D03.01), further work is required to refine the algorithms already tested or, if necessary, to try additional techniques that can lead to more accurate results. This further work will be carried out in the next phases of this project.