D5.2 WATSON Platform design and alignment to vision

This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No. 777439

Watson

Innovation analytics

AUTHOR NAME	**Organization**
ANNA PALAIOLOGK	EXODUS
ZISIS ARAMPATZIS	EXODUS
SPIROS TROUGKAKOS	EXODUS
PAOLA DE PASCALI	KAPITALISE
MARTYNAS KISKIS	INVENTYA
RALF MARTIN	IMPERIAL COLLEGE OF SCIENCE TECHNOLOGY AND MEDICINE

REVIEWER NAME	**Organization**
STEPHEN BOURNER	KAPITALISE
AUŠRA LINGYTĖ	INVENTYA

VERSION	**Date**	**Modifications**
0.1	02/03/2018	Table of Contents
0.2	15/03/2018	Architecture updated
0.3	30/03/2018	Data described
0.4	30/04/2018	User scenarios and functionality updated
1.0	09/05/2018	Final updates and comments integration

This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No. 777439

Watson
Innovation analytics

Executive Summary

The WATSON project aims at developing a framework methodology for studying the impact of R&D tax credits and incentives on SME-led innovation in Europe implemented in a user-friendly ICT platform. The main users of the WATSON ICT platform will be governing bodies and private investors, as well as SMEs. Deliverable D5.1 reports on the user requirements of the WATSON platform. User needs have been gathered from consortium partners with respect to five dimensions:

- Data: Which (type of) data should the tool be able to process? How should WATSON relate to other tools and data sources? What are the privacy concerns involved?
- Bottlenecks of existing solutions: Which are the existing solutions and how do they serve the needs of target users?
- User needs & functionality: How will the user interact with the tool and what are the requirements regarding capabilities?
- On technical issues: What should be the architecture of such a tool?
- Standardisation & GDPR: What are the standardisation activities planned and how will WATSON comply with GDPR?

Deliverable D5.2 will continuously be updated through interaction with the other work packages and will serve as a guiding document throughout the project lifetime for the technical implementation.

This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No. 777439

This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No. 777439

Contents

Executive Summary 3

1 Introduction 7

2 User requirements 8

2.1 Description of Data 8

2.2 Bottlenecks when using existing solutions 12

2.3 User identified needs 12

2.4 Foreseen functionality 14

2.5 Standardization and privacy concerns 15

3 Watson platform architecture 17

3.1 Service oriented architecture (SOA) 18

3.2 Services 18

4 Watson standardisation and GDPR issues 22

5 Conclusions 23

This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No. 777439

Watson

Innovation analytics

This project has received funding from the European Union’s Horizon 2020 research & innovation programme under grant agreement No. 777439

1 Introduction

The general objective of the WATSON project is to develop and to demonstrate a framework methodology for studying the impact of R&D tax credits and incentives on SME-led innovation in Europe. The framework and results of the project will be integrated into an ICT platform that performs data analytics on innovation funding, identifying gaps and allowing the consortium partners to advise stakeholders (public governing bodies, SMEs, private investors) on the management of innovation and funding schemes.

The final result of the project is thus twofold:

- The delivery of a framework methodology to study the impact of R&D funding on SME-led innovation, which can be extrapolated to study further funding measures other than R&D tax credits and incentives. The results will be made publicly available, delivering tools to identify funding gaps in SME segments with high impact potential, as well as proposing new measures for better targeting R&D tax credits and incentives to maximize impact of innovative SMEs.
- The delivery of an ICT platform that integrates the results of the project to provide innovation analytics services. The platform will be used by the consortium partners to commercialise advisory and consulting services on innovation funding for customers from the stakeholder groups and thus take to market solutions that tackle the current shortcomings of innovation funding.

This deliverable mainly focuses on the user requirements for the second objective, the development of an ICT platform that provides interested stakeholders and users with a user-friendly way to access the results of the first objective. Through this platform, the analysis of SME innovation impact will be correlated with the analysis of R&D tax credit and incentive funding, in order to identify funding gaps and opportunities. This deliverable will be the main input for the first release of the platform (Milestone 5 of WATSON project). All input has been gathered from consortium partners based on their knowledge of the ecosystem and their interactions with end-users. After the first prototype is released, the consortium will seek direct feedback from early-adopters, through the relevant exploitation activities.

This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No. 777439

2 User requirements

2.1 Description of Data

Various data are considered in WATSON supporting required outcomes. These data will be adapted and integrated in the WATSON system, mainly via offering the starting points for the generation of features to be analysed in combination in the Cloud.

Data	Description/Type	Foreseen use in WATSON
European Patent Registry database	• Data entries are described in Rule 143 of the European Patent Registry. • number of the European patent application; • date of filing of the application; • title of the invention; • classification symbols assigned to the application; • the Contracting States designated; • particulars of the applicant for, or proprietor of, the patent; • family name, given names and address of the inventor designated by the applicant for, or proprietor of, the patent; • particulars of the representative of the applicant for, or proprietor of, the patent; • priority data (date, State and file number of the previous application); • date of publication of the application and, where appropriate, date of the separate publication of the European search report; • date of filing of the request for examination; • date on which the application is refused, withdrawn or deemed to be withdrawn;	• Should correlate with the EU CIS database; • Database has been extensively used with multiple innovation studies.

	• date of publication of the mention of the grant of the European patent; • date of lapse of the European patent in a Contracting State during the opposition period and, where appropriate, pending a final decision on opposition; • date of filing opposition; • date and purport of the decision on opposition; • dates of stay and resumption of proceedings; • dates of interruption and resumption of proceedings; • date of re-establishment of rights; • the filing of a request for conversion; • rights and transfer of such rights relating to an application or a European patent where these Implementing Regulations provide that they shall be recorded; • date and purport of the decision on the request for limitation or revocation of the European patent; • date and purport of the decision of the Enlarged Board of Appeal on the petition for review.	
Eurostat Community Innovation Survey (CIS) database	• New or significantly improved products; • Introduction of new processes; • New distribution or logistics methods; • Characteristics of innovation activity at the enterprise level.	• Could correlate with the European Patent Registry database; • Data could be used for direct visualization; • Could work on both the cloud & local level.
WATSON proprietary innovation survey	TBA	TBA

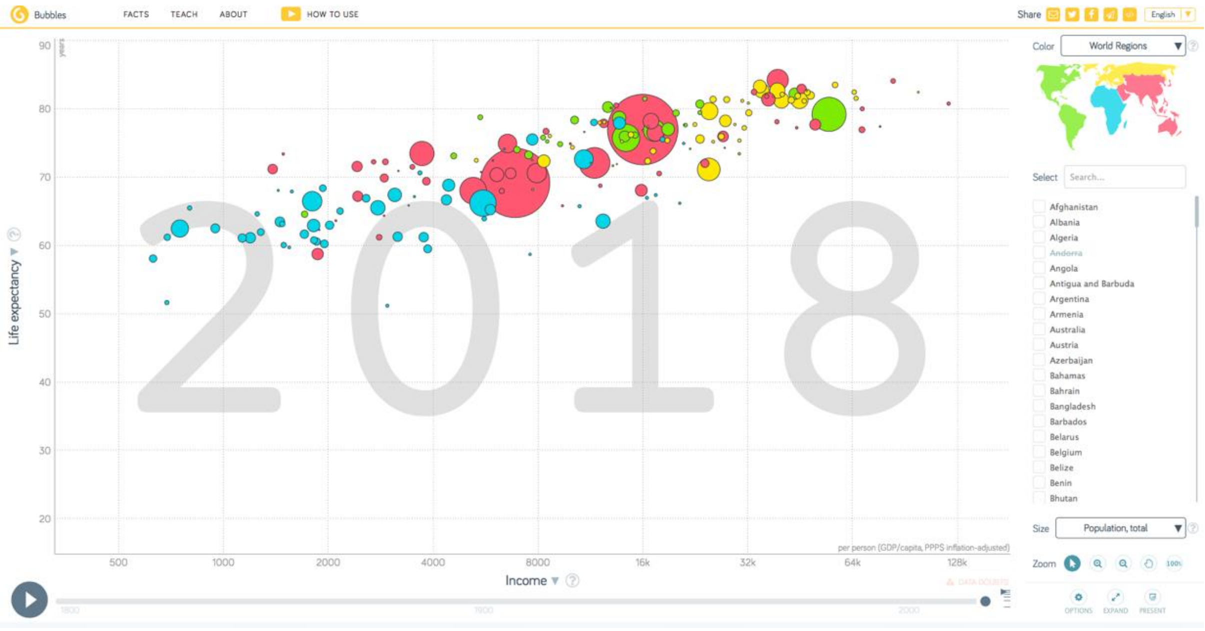
Innovation counts (Angel List)	• Description of the innovation; • External funding amount (self-reported); • Industry; • Key team members.	• Data could be used for calculation – innovation counts methodology. • Data potentially could be used to replicate Innovation Counts methodology; • Data available through open APIs.
Innovation counts (Crunch Base)	• Description of the innovation; • External funding amount (self-reported); • Industry; • Key team members.	
Government R&D policy looking at 10 countries (Denmark, France, Greece, Ireland, Italy, Lithuania, Netherlands, Romania, Spain and UK). **WP.3.1**	Policy and procedure data with summary for each country involved: • Type of incentives in R&D tax credits (e.g. incremental scheme vs volume-based scheme). • Eligibility- Eligible Costs (e.g. Eligible expenses and eligible sectors). Process of receiving R&D tax credits.	Data could be used for direct visualization.
Innovation strategies looking at 10 countries (Denmark, France, Greece, Ireland, Italy, Lithuania, Netherlands, Romania, Spain and UK). **WP 3.1**	• Economic overview (e.g. GDP, number of SMEs in the country). • Innovation strategies (e.g. gross domestic expenditures on R&D – GERD and key sectors).	Data could be used for direct visualization.
Qualitative research Watson proprietary. **WP3.2**	Primary data (collected though qualitative open-ended questions from SME representative organisations, government agencies and interested stakeholders) in order to collect:	Data could be used for direct visualization.

	• Knowledge based on Research and Development tax incentives. • Barriers to innovation • Stakeholder Aspiration Strategy content of the organizations in order to promote R&D tax incentives and support for SMEs.	
Quantitative research Watson proprietary. **WP3.3** (still to be defined)	Following Inventya's new segmentation, measuring the impact of R&D tax credits and incentives on SME innovation through a questionnaire survey. Quantitative variables specifically proxy variables such as: OUTPUT OF INNOVATION • Number of patents • Number of new products/services • Marketing innovation (work in progress)	Data could be used for direct visualization. Could correlate with the CIS.
Innovations	Database containing characteristics of all global innovations (i.e. patent families) as available from PATSTAT	Used in conjunction with citation link data (and possibly company level and inventor data) to produce patent rank
Citations	Database containing citation vectors (i.e. innovation A cites innovation B)	Used in conjunction with innovations data to produce patent rank
Company data	Database public Company data	
Inventor data	Database containing public inventor data	
Inventor innovation link	Database linking innovations to inventors (as inventors do many inventions and inventions are often connected to multiple inventors it's efficient to store this info in separate database	
Inventor company link	Database linking	

2.2 Bottlenecks when using existing solutions

After analysing the market for alternative tools/platforms to perform innovation segmentation, we identified the following bottlenecks:

- No existing solution provides information on Patent Rank to guide policy
- No existing solution may be used to slice and dice innovation in a visual way.
- Most current innovation research and case studies have been mostly presented in an academic format, thus do not provide any degree of interactivity.
- The closest Cloud-based tool for economic analysis to Watson could be Gap minder. However, the Gap minder does not provide any variables for innovation research and is limited to common macro-economic variables such as life expectancy, income level, average family size and other.



Color: World Regions

Size: Population, total

Figure 1 Gap minder is a Cloud-based tool for economic analysis, however it does not include any variables necessary for innovation studies.

2.3 User identified needs

Scenarios are described below for two different types of user of the platform (governing bodies and private investors) and what they may use WATSON for.

This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No. 777439

Scenario 1: A private investor uses WATSON to validate their assumptions about new investment opportunities

A private investor could use WATSON to validate their assumptions about emerging industries that they consider investing in.

The investor might want to understand the following:

- Has there been an increase in patenting activity in this particular industry segment within a specific time period?
- Did companies in the specific industry segment raise more capital than other innovating companies?
- Did a particular industry segment produce more innovative products or services than other segments (compared with the official EU Community Innovation Survey data);
- Is a particular innovative segment concentrated in a specific geographic area? If yes, where are these areas?
- What are the other innovative segments in a particular geographic area? Are there any new innovative clusters emerging?

Scenario 2: A cross-national comparison of R&D tax incentives

A researcher/policy-maker may need to quickly open specific paragraphs and sections in the regulation of different countries to be able to quickly compare them. The platform could provide a summary table for easy comparability of R&D tax incentives around Europe (e.g. an image with a geographical map of Europe where it is possible to click on the country and read all the information about R&D tax incentives there).

Scenario 3: A private investor could use WATSON to look at different R&D policies around Europe

The investor might need to understand the following:

- Using the summary policy table, a private investor could quickly know where (in which country) it is more convenient to invest in and in which sectors.
- Where the R&D tax incentives are more generous?
- What are the most advanced/innovative sectors in this country?

Scenario 4: To visualise the strengths and weaknesses of R&D tax incentives in the selected countries and the barriers of innovations.

Following the results of our two empirical works (WP3.2 and WP3.3) respectively the qualitative and quantitate surveys, the users (researchers, policy makers and the private investors) might need to understand the following:

- Strengths and weaknesses of R&D tax incentives in the selected countries.

This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No. 777439

- The obstacles to innovation for each country involved in our research (e.g. lack of a shared vision, purpose and/or strategy, lack of time, resources and staff, bureaucratic barriers to claim back R&D tax incentives).
- To quantify the impact of R&D tax incentives.
- Scenarios for Patent Rank
- A civil servant, a consultant advising government, wants to compute Patent Rank Figures that are specific to their area of interest; e.g. for a region there is an interest in understanding the contribution of small vs large firms from a specific technology area to knowledge spillovers relevant to the region. If spillovers from small firms are particularly large this can justify a dedicated support programme for R&D from such firms. The PatentRank app will allow the user to set the parameters requirement (Region R, Technology area X, Firm type J) and will then compute the relevant results. It will also provide various visualization tools for the results

2.4 Foreseen functionality

The scenarios described above involve the functionality listed in the table below.

Functionality	**Description**
Show policies per country	The user should be able to select a country (e.g. UK) and see all of the relevant R&D policies.
Compare countries	The user should be able to open two windows to compare the policies of the two selected countries
Filter R&D policies	Policies should be categorised so that the user can easily and quickly find the sector he or she is interested in.
Map	There should be a map so that users can click on the nation to open its list of policies
Show summary	Each country will have a section highlighting a summary containing: strengths and weaknesses of R&D tax incentives, plus the effect of R&D tax incentives in terms of output of innovation (e.g. In UK 50% of companies that received tax incentives have doubled their turnover as compared to the others that did not receive it)
Graphical interface to select Patent Rank parameters	Patent Rank can be calculated, conditional on a variety of parameters and assumptions. To a large extent this depends on the users' interests and requirements. The app will have a user-friendly web interface for users to define these parameters.

Computation of Patent Rank	The server will compute Patent Ranks. This requires a separate process that can also be accessed programmatically via an API. This would allow more advanced users, as well as members of the project team, to access the process programmatically to compute figures that are then fed back.
Patent Rank exporting suite	After Patent Rank calculation, users can specify how the output is being fed back to them (e.g. download as spreadsheet, sent to visualisation engine)
Map function: display most innovative industries per country/ region	The user could view what are the most innovative industries per region or country. They also could choose how they want to measure innovation: number of patents, % of R&D personnel, innovation counts.
Map function: display key technologies used to create innovations per country/ region	The users could view what are the key underlying technologies used to produce innovation in specific regions/ countries. E.g. Cloud company, machine learning, advanced materials, 3D printing etc.
Segmentation function: display key underlying technologies used to produce innovations in specific industries	The user could view what are currently key technologies used to come up with innovative products & services (regardless of the region/country). E.g. machine learning in IT industry, advanced materials in automotive industry, 3D printing in apparel industry.

2.5 Standardization and privacy concerns

There should not be any privacy concerns using anonymized data from the European Patent registry and Community Innovation Survey (Eurostat). This data has been extensively used by numerous researchers to analyse innovative activities in the European Union, both at the country and enterprise levels.

Data that the WATSON project will collect through primary research may be subject to privacy issues. The project should only use anonymized data excluding any company details that could be used to identify a particular company. We would like to avoid situations where the WATSON tool is used for due diligence on any particular company, especially when such company might be actively seeking external funding. Hence, the tool should not be used to artificially boost a company's public image, nor diminish it, as it may cause legal issues both for investors and investment-seeking companies.

Innovation analysis should be done at the industry level with the intention to help policy makers, investors and entrepreneurs understand hidden traits of innovative SMEs and industries without revealing any personal information of the specific companies.

Hence, if the project uses self-reported data from open databases such as the Crunch base, personal information such as the founder's name, social media accounts, company name, and product titles should be completely anonymized.

3 Watson platform architecture

For the Watson platform we suggest that the most suitable architecture is a service-oriented architecture. The purpose of this approach is to design a highly decoupled architecture which is extensible and easily maintainable. The benefits of this architecture are many and varied. Most important benefits are: technology heterogeneity, fault tolerant, scaling, deployment and composability.

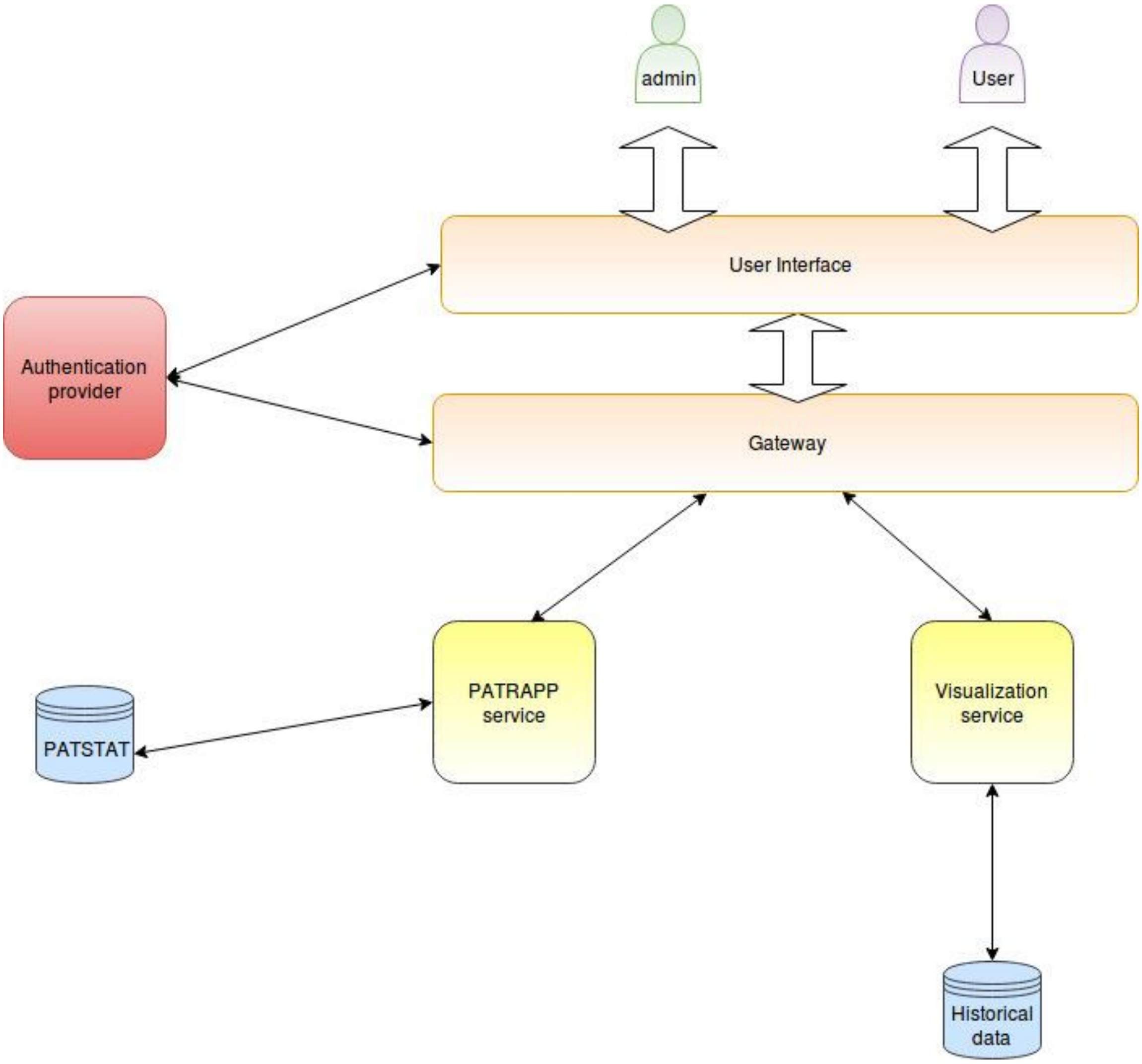


Figure 2: Watson architecture

3.1 Service oriented architecture (SOA)

Below we can see a brief description of the benefits of a SOA.

Technology heterogeneity

With a system composed of multiple collaborating services, we can decide to use different technologies inside each one. This allows us to pick the right tool for each job.

Fault tolerant

A key concept in resilience engineering is the bulkhead. If one component of a system fails, but the failure doesn't cascade, you can isolate the problem and the rest of the system can carry on working. With this architecture we can build systems that handle the total failure of services.

Scaling

With small services, we can just scale those services that needed scaling, allowing us to run other parts of system on smaller, less powerful hardware.

Deployment

In SOA, we can make a change to a single service and deploy it independently of the rest of the system. This allows us to get our code deployed faster. If a problem does occur, it can be isolated quickly to an individual service, making fast rollback easy to achieve. It also means we can get our new functionality faster.

Composability

One of the key promises of distributed systems and service-oriented architectures is that we open up opportunities for reuse of functionality. We allow our functionality to be consumed in different ways for different purposes.

3.2 Services

The Watson architecture will have few basic services (Authentication, visualization, patent rank algorithm) and a Gateway that will help for the routing of the requests for user authentication through authentication service and for services integration.

User Interface

The user interface will provide two panels, one for the administrators of the platform and one for the simple users.

The user interface should provide the end user with the possibility to apply filters to the data that will be used as input to the patent rank algorithm and also to apply filters to the visualization of the parameters the end user wants to see.

Authentication server

The platform will be secured by an Authorization server which is based on OpenID Connect and JWT. OpenID Connect provides authentication flows that are based on redirection. Upon successful authentication, an access token is issued. This access token is subsequently being used (over HTTPS) from every client when making requests to the platform. Backend services need to validate that the access token is valid. In order to have every service aware of the security infrastructure in-place (token decoding, authorization server etc.), we will use a middleware service that sits in front of every exposed backend service. This middleware service is, among other things, responsible for validating the access token, resolving the corresponding user identity and roles, and forwarding an amended request to the target backend service that will include this security context (user identity and roles).

PATRAPP service

This service will provide a REST API to patent rank algorithm. This service will make all the necessary pre-processing of the data so as to be in the right structure for the patent rank algorithm. This service will also retrieve the data from PATSTAT. This service should allow users to compute the patent rank using the latest data available and compute the algorithm with different attributes of innovation. The patent rank algorithm should also be available in different versions. For this reason patent rank microservice will provide different versions of REST APIs, for different versions of the algorithm.

Innovation data

The raw data needed for patent rank algorithm will be retrieved from PATSTAT, a comprehensive database of patent applications to all global patent authorities, along with various pieces of auxiliary information. In particular, we require data on patent citations as well as information on patent families; i.e. a family are the multiple patent documents connected with the same underlying innovation; i.e. many innovations are patented multiple times in different jurisdictions. The data on patent families identifies which patent documents refer to the same underlying innovation (Note that the same innovation can also have multiple patent documents in a single jurisdiction because different countries have different rules regarding what can be considered one patent; e.g. the Japanese patent office has a tendency to split between several patent documents what is being combined in just one document in many other jurisdictions). The main unit of analysis for the purpose of patent rank is innovation; i.e. using the families file we will aggregate everything to the level of an innovation.

Hence, after processing of the raw data we would end of up with two main data files which can be seen as the nodes and edges of a network

- Innovation data (Nodes)

  For every innovation we record (at a minimum, we might expand this list)

  - Unique ID
  - Application year
  - IPC classification (technology type)
  - Patent holder ID (can be multiple patent holders)
  - Patent holder geolocation (latitude, longitude)
  - Patent holder country
  - Innovator ID (can be multiple)
  - Innovator geolocation
  - Innovator country
  - Granted year (missing if not granted)
- Citation data (edges)

  Which citation is cited by whom?

  At a minimum this consists of the following variables.

  - cited_ID
  - citing_ID
  - Total number of cites by a citing innovation

Visualization service (visualize engine)

This service will provide a REST API that will create dynamic charts and dashboards that will be easily embedded to an HTML page. This service should be able to visualize historical data (data from questionnaires and patent counts).

Functional requirements:

Below are defined some basically functional requirements.

- static charts
- interactive charts
- custom charts
- support multiple data sources
- widgets
- dashboards

Static charts

The module will be able to provide static charts (images) for a variety of types such as bar charts, pie charts, scatter plots, timeseries etc.

Interactive charts

The interactive charts will be used from web browsers. Interactive charts will be easily embedded in html pages. All static charts will be also available in interactive mode. Some features of interactive charts will be the zoom in, zoom out, hover tools, selections tools, reset chart and save the interactive chart into an image.

Custom charts

A variety of chart customizations will be available to the end users such as colour, shape, width etc.

Support multiple data sources

The microservice will support a variety of inputs such as csv, json Excel, etc.

Widgets

A variety of widgets will be available like buttons, checkbox buttons, dropdown menus, multiselect, radio button, selection button, sliders etc. The end user will be able to change the visualized data.

Dashboards

The user will be able to define his/her own custom dashboards, with as many charts and widgets he wants.

Gateway

This service will provide a REST API. This service is responsible for checking if a user is authorized and the role of the user, and to forward the requests to the right backend services. This service will be the entry point to the backend Watson platform and integrate all Watson services.

4 Watson standardisation and GDPR issues

Today, there might be several data standardisation issues that must be solved:

- The acceleration of data volumes, velocity (if open APIs are used to fetch data in real time), variety and veracity may overwhelm the data management architecture;
- Given multiple data sources redundancies may occur. These redundancies would be counterproductive to insight generation;
- Data storage costs – the project should attempt to reduce the data storage costs by reducing the amount of data stored or the granularity of data.
- Individualization of user interactions presents a scalability issue.

In terms of the GDPR issues, again data sourced from public sources such as the European patent registry and the Community Innovation Survey will comply with the European GDPR regulations without any alterations from the WATSON project team.

The WATSON survey data, along with the data retrieved from self-report databases will need to be anonymized in compliance with the GDPR rules.

Recital 26 of the GDPR defines anonymized data as “data rendered anonymous in such a way that the data subject is not or no longer identifiable.” Although circular, this definition emphasizes that anonymized data must be stripped of any identifiable information, making it impossible to derive insights on a discreet individual, even by the party that is responsible for the anonymization.

Additionally, WATSON data pseudonymization could be possible. Article 4(5) of the GDPR defines pseudonymization as “the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information.”

By rendering data pseudonymous, WATSON could benefit from more relaxed standards under the GDPR. For instance, Article 6(4)(e) permits the processing of pseudonymized data for uses beyond the purpose for which the data was originally collected. Additionally, the GDPR envisions the possibility that pseudonymization will take on a key role in demonstrating compliance under the GDPR.

Both Recital 78 and Article 25 list pseudonymization as a method to show GDPR compliance with requirements such as Privacy by Design.

5 Conclusions

This document presents the required analysis for developing a platform that will allow governing bodies and private investors to get an insight and take decisions based on the situation with regards to the impact of R&D tax credits and incentives on SME-led innovation in Europe. The requirements cover both the perspective of the data processors and the final end-users. In this deliverable we have analysed the general project scenario and split it into four sub-scenarios regarding: validation of investment opportunities, cross-national comparison of R&D tax incentives and R&D policies, and visualisation of barriers to innovation.

We have described in detail how the users will interact with the system in the user functionality table, making up 11 functional and non-functional requirements.

The technical requirements are directly related to the use such a platform may have in the future and address functionalities such as: integration of external data sources, stand-alone services, data visualisation and user interaction with the data, authentication, and data privacy protection. The main benefits of the chosen architecture are: technology heterogeneity, fault tolerant, scaling, deployment and composability.

Finally, WATSON is monitoring all relevant standards and data privacy laws in Europe, which will ensure the compliance of its solution with the latest updates.