3.7 COMPARED - Text mining solution to support the evaluation process of research grant applications (2018.07)

3.7.1 IDENTIFICATION OF THE ACTION

Service in charge	JRC.I.3
Associated Services	RTD

3.7.2 EXECUTIVE SUMMARY

Public funding agencies are investing billions of Euros in research and innovation (R&I) projects every year. Funding mechanisms can be improved to reach higher funding efficiency e.g. by aiming at the reduction of unnecessary duplication or overlaps between research proposals, increasing the quality of incoming proposals and decreasing the number of submitted R&I projects. There is also no doubt that the process of evaluating research proposals should be based as much as possible on scientific evidence. One way funding agencies could work towards this is by facilitating the sharing to other agencies of data related to public funding of research in Europe. But not all funding agencies have sufficient expertise in data analytics to act on this issue and the European context, with many funding mechanisms at regional, national, or European levels, does not help. This diversity of funding mechanisms is an asset but also a burden as it makes connecting funding schemes together difficult.

Through the development of a semantic similarity platform that would select documents relevant to the evaluation process, COMPARED aims at supporting evidence-based decision-making in the field of public funding of R&I. The project aims to achieve data interoperability but not interoperability of IT systems. Indeed, overall interoperability does not hinge on data availability of funded research alone and actually depends on systems design, processes and rules, which are context specific and therefore legitimately localised. By giving funding agencies, applicants and other stakeholders access to a semantic platform for the assessment of research proposals, the project aims to contribute at reducing unnecessary research duplication, reducing scientific overlap between funded projects, and at increasing the quality of R&I proposals while reducing the number of incoming proposals. Recent publications have identified these issues as key to maximise the impact of publicly-funded R&I[74,75,76]. This was also confirmed in a recent report by an independent high-level group recommending the European Commission to align national and EU R&I investment schemes, establish synergies with other funding programmes in Europe, and increase the impact of publicly-funded research in Europe[77].

Applicants to publicly-funded research programmes could also benefit from means to verify how similar their proposal is to funded R&I projects and other documents (e.g. scientific publications or patents). This would help applicants submit more original projects or help justify why research has to be duplicated, and will contribute to increasing the quality of research proposals entering the evaluation process at public funding agencies. Another benefit of giving access to grant data to applicants would be to reduce the incoming number of grant applications for funding agencies, as applicants would receive indications on similar projects already funded. This reduction of incoming

[74] *Concentrating on the Fall of Labor Share; CEP Discussion Paper No. 1476; Grell, Kevin Berg – Marom, Dan – Swart, Richard (2015): Crowdfunding, The Corporate Era, Elliott and Thompson, London, 218 p.*

[75] *Funding agencies urged to check for duplicate grants, Nature, January 2013, volume 493.*

[76] *The Economic Rationale for Public R&I Funding and its Impact, European Commission DG Research & Innovation, ISBN: 978-92-79-65270-7*

[77] *"Lab-Fab-App, investing in the European future we want", Report of the independent high level group on maximising the impact of EU research & innovation programmes. European Commission DG Research & Innovation, ISBN: 978-92-79-70069-9*

proposals would have be a significant added value for funding agencies as it could reduce operational costs related to grant evaluation. In addition, as most of R&I today is privately funded, making some parts of COMPARED publically accessible would also allow private actors of R&I (companies, investment firms) to use the platform to reduce duplication in R&I investments and overlap between research projects.

The deliverables of the COMPARED project will consist of a pilot web-based platform, the first version of the database containing grants data and the system for collecting data, and a set of recommendations for possible further extension and full deployment of the system. From the technical point of view, preliminary tests have been performed to assess the technical feasibility of such a semantic retrieval of documents, based on the text of an incoming R&I proposal. The results of these tests were positive. The pilot platform that will be developed during the project will be based on user requirements provided by experts involved in the project and by the advisory board. This key input will be collected at the beginning of the project to drive the design of the platform. This will maximise impact on the evaluation process and help customise the platform with relevant features and visualisations. During the pilot phase, legal issues related to data will be explored and various solutions for translation of research proposals into English will be tested. To prepare for a possible wide dissemination of the platform, contacts will also be made with additional member state funding agencies and associations. A panel of experts in grants evaluation will accompany the project. This panel will review the work accomplished and set a list of recommendations for further development and deployment.

The Joint Research Centre of the European Commission has a solid expertise in text and data mining in which it is active for more than 15 years[78]. The present project will be located in the Text Mining Competence Centre recently launched by JRC to serve the Commission with text mining solutions.

3.7.3 OBJECTIVES

The overall objective is to confirm the feasibility and usefulness of a semantic platform for the evaluation of research proposals. Specific objectives are:

1. Develop a pilot web application that evaluators of R&I proposals can use to obtain similar documents relevant to the evaluation process. This platform would provide additional information useful for grant assessment but does not aim at replacing existing evaluation processes used by agencies.
2. Develop the first version of the database containing the corpus of data needed for the semantic comparison of research proposals and of the system to collect data. Data on research grants will be coming from European funding bodies (e.g. Commission or Eureka) and from national funding agencies. Additional data related to patents and to scientific literature will be considered as well.
3. Reach out to stakeholders and create a community of potential users to drive the development of the COMPARED platform.

3.7.4 SCOPE

This project aims to provide for the design, development, implementation, and operation of a semantic similarity pilot platform to support the process of evaluating research proposals. The end product will be a pilot web-based application, where users can retrieve documents semantically similar to the

[78] *Check http://emm.newsbrief.eu and http://www.timanalytics.eu for concrete examples of IT solutions.*

proposal they are evaluating at the time. The project will also deliver a recommendation report from a group of experts, confirming or disproving the usefulness of such a platform and a possible scale-up. It should be noted that the semantic similarity platform does not aim to replace IT systems used to perform evaluation of proposals, neither does it aim to harmonise evaluation processes for research proposals throughout Europe or data standards. Rather, it aims at complementing processes operated in Member States by creating a bridge between evaluation processes.

3.7.5 ACTION PRIORITY

3.7.5.1 Contribution to the interoperability landscape

Question	Answer
How does the proposal contribute to improving interoperability among public administrations and with their citizens and businesses across borders or policy sectors in Europe? *In particular, how does it contribute to the implementation of:* • *the new European Interoperability Framework (EIF),* • *the Interoperability Action Plan and/or* • *the Connecting European Facility (CEF) Telecom guidelines* • *any other EU policy/initiative having interoperability*	The project will ignite data interoperability in a field where a real need for more cross-border collaboration exists, but for which there are no IT solutions yet. Some initiatives like the Lead Agency Model offer models for cross-border collaboration but there exists today no means to compare R&I grants at a European scale. The first benefit of the project will be to establish data interoperability between funding agencies in different member states. This will be done with minimum disturbance to processes operated today by funding agencies: there will be no direct impact of the COMPARED platform on IT systems operated by public funding agencies. In addition the web application will be accessible through simple url links. The current project is in line with 2 ERA priorities[79] and with a recent report by an independent high-level group delivered to DG Research and innovation, which encourages the European Commission to align national and EU R&I investment schemes, to establish synergies with other funding programmes in Europe, and to increase the impact of publicly funded research in Europe[80]. The project will also contribute to opening up access to grants data, which is common practice e.g. in the US and the UK. Opening access to grants data, however, can only be of real value if there is a single point of access to the data. Through the

[79] "More effective national research systems that include increased competition within national borders and sustained investment in research" and "Transnational cooperation and competition which define and implement common research agendas on challenges, raise quality through Europe-wide open competition, and construct and run key research infrastructures on a pan-European basis".

[80] "*Lab-Fab-App, investing in the European future we want*", Report of the independent high level group on maximising the impact of EU research & innovation programmes. European Commission, DG Research & Innovation, ISBN: 978-92-79-70069-9

requirements?	COMPARED platform data that are today not available would be made so in a common format. Openness will also apply to the project itself, which will involve real users from design to testing and validation. Dissemination and access to data will be royalty-free, but restricted to non-profit activities.
Does the proposal fulfil an interoperability need for which no other alternative action/solution is available?	There are today no IT solutions for addressing the lack of informed decision-making, when it comes to the evaluation of research project proposals. Some local solutions exist, however they cannot work in isolation. The real issue is related to the fragmentation of the funding mechanisms in Europe and the difficulty to gather the relevant corpus of data, combined to the possibility for project applicants, organised in consortia, to submit grant proposals across borders. An EU-wide approach including grant data from FP and ERC programmes would guarantee a meaningful volume of data.

3.7.5.2 Cross-sector

Question	Answer
*Will the proposal, **once completed** be useful, from the interoperability point of view and utilised in two (2) or more EU policy sectors? Detail your answer for each of the concerned sectors.*	Should the project be successful, it could contribute to enhanced evidence-based decision making and provide some elements for more cross-border collaborations in that field. Data interoperability (and not system interoperability) would be achieved through collecting data from the different funding mechanisms in Member States via the COMPARED platform. Funding of research projects by public organisations is a cross-sector activity. Once implemented, the IT solution proposed here will contribute to more informed decision-to-fund in various policy fields like energy, environment, ICT, health, transport, and many more.

3.7.5.3 Cross-border

Question	Answer
*Will the proposal, **once completed,** be useful from the interoperability point of view and used by public*	1) Administration to Administration. Once completed, the platform will be used by as many funding agencies of Member states as possible, ideally by agencies in all Member States, as well as in other countries. The project will establish close interaction with National funding agencies and with Science Europe (gathering funding agencies from many Members States), with

administrations of three (3) or more EU Members States? Detail your answer for each of the concerned Member State.	the goal to involve the final users as soon as possible in the project. We will also aim for a maximum of these funding agencies to contribute to COMPARED with data about grants. For funding agencies that have strong expertise in evidence-based evaluation of research proposals, the main advantage in using the platform will be mainly to obtain information about research projects funded in other Member States. In addition to this, funding agencies less advanced in evidence-based decision-making will also be to share best practices in the evaluation of research proposals and of their impact. 2) Administration to citizens & administration to business. Parts of COMPARED will be publically accessible allowing applicants to build more innovative proposals and investment funds or companies to better evaluate requests for R&I funding.

3.7.5.4 Urgency

Question	Answer
Is your action urgent? Is its implementation foreseen in an EU policy as priority, or in EU legislation?	Although there is as such no urgency, evidence-based decision-making in the funding of R&I projects by public agencies is critically needed. Evaluators of grants have no means of knowing if a particular research project has already been funded elsewhere, or if the research has already been performed. Experts use their vast knowledge and experience to evaluate the originality of projects, but there are no actual systematic prior art searches being performed as part of the evaluation process. Knowing more about the past will help evaluators to assess the quality of research proposals and justify their decision on more factual elements. Ideally the platform should be fully operational for the start of FP9 in 2020.
How does the ISA^2 scope and financial capacity better fit for the implementation of the proposal as opposed to other identified and currently available sources?	This project fits with the ISA^2 interoperability goals. There are no other identified available sources of funding for this project.

3.7.5.5 Reusability of action’s outputs

Name of reusable solution to be produced (for new proposals) or produced (for existing actions)	**COMPARED platform**
Description	The platform will be accessed through a web application and will therefore be re-usable by any additional funding agency or other entity wishing to use it, subject to certain limitations related to ownership of data. No personal data will be needed for the project.

Reference	
Target release date / Status	Re-use is part of the project. Platform accessible and available as the project evolves and on request.
Critical part of target user base	Funding agencies.

Name of reusable solution to be produced (for new proposals) or produced (for existing actions)	**COMPARED data**
Description	To the extent that is possible, the dataset on which the platform will rely will be made available to funding agencies and possibly other stakeholders, with the condition that the data can be exclusively re-used for non-profit activities.
Reference	
Target release date / Status	Re-use is part of the project. Data will be made available from the onset, depending on specific legal or data protection issues.
Critical part of target user base	Funding agencies, scholars in the field of scientometrics, economics, innovation and research management.

Name of reusable solution to be produced (for new proposals) or produced (for existing actions)	**COMPARED code**
Description	Finally, the JRC code will be made available through licensing schemes without royalty compensations.
Reference	
Target release date / Status	Re-use is part of the project. JRC Code accessible will be made available as much as possible as the project evolves and on requests.
Critical part of target user base	Developers of text mining solutions.

3.7.5.6 Level of reuse of existing solutions

Question	Answer
Does the proposal intend to make use of any ISA^2, ISA or other relevant interoperability solution(s)? Which ones?	EUPL whenever possible. PM^2. Possibly DCAT-AP, but this will have to be analysed further.

3.7.5.7 Interlinked

Question	Answer
Does the proposal directly contribute to at least one of the Union's high political priorities such as the DSM? If yes, which ones? What is the level of contribution?	Contribution to "Boosting competitiveness through interoperability and standardisation". Less duplication of research means more original research funded, hence some impact on competitiveness.

3.7.6 PROBLEM STATEMENT

The problem of	The difficulty to perform prior art search before evaluation of grant proposals
affects	The amount of evidence useful to assess whether a particular proposal should be funded or not.
the impact of which is	No evidence-based decision-to-fund.
a successful solution would be	Provide a semantic similarity platform that will automatically deliver to the evaluator a set of documents similar to the proposal under evaluation.

The problem of	Variety of local IT legacy systems.
affects	Technical interoperability
the impact of which is	Difficult to link systems together and exchange data
a successful solution would be	A centralised repository for data on grants, accessible through a semantic web application easy to integrate or embed in existing processes, with data exchange using RSS format and specific semantics and syntactic.

The problem of	Heavy workload related to processing of research projects.
affects	Efficiency of funding agencies.
the impact of which is	Reduced capacity for sound decisions and to accompany applicants.
a successful solution would be	Give access to a semantic platform to applicants may help in reducing the number of proposals for funding.

The problem of	Limited access of applicants to data on previously funded research projects or to other relevant scientific documents.
affects	The quality and novelty of research projects.
the impact of which is	Proposals entering the evaluation process are of lower quality and novelty than expected, which has an impact on competitiveness and innovation potential.
a successful solution would be	Give access to a semantic platform to applicants may help in increasing the quality and novelty of proposals for funding.

The problem of	High fragmentation of many funding schemes operating in Europe.
affects	Cross-border collaboration, which is low, and exchange of data, which is rare, and therefore the capacity to detect multiple funding of research and overlap of research grants.
the impact of which is	Lack of novelty in proposals, overlap between research grants, and duplication of research.
a successful solution would be	Give access through a semantic platform to a corpus of data on research projects funded in EU Member States, at EU level, or outside.

3.7.7 IMPACT OF THE ACTION

3.7.7.1 Main impact list

Impact	Why will this impact occur?	By when?	Beneficiaries
(+) Savings in money	Detection of overlaps in research projects (scientific and financial) and subsequent reduction in overlaps and research duplication.	Q1 2020	Funding agencies (Member States and others)
(+) More innovation	More innovative R&I projects.	Q1 2020	Member States
(+) Interoperability	There is no interoperability in this field.	Q1 2020	Funding agencies (MS and others)
(-) Integration or usage cost	Any new tool is associated to some costs: training, integration in IT, licensing, data exchange… But costs will be limited, as the platform will consist in a web application. Impact on agencies will be minimal, in particular because the use of the platform will have no impact on the IT systems in operation locally.	Q1 2020	Funding agencies (MS and others)
(+) More evidence-based funding decisions	Evaluators would have access to prior art documents retrieved through a semantic process.	Q1 2020	Funding agencies (MS and others)
(+) Open access to data on research grants	Catalyse open access to grant data and provide a central access point	Q1 2020	All innovation stakeholders.

3.7.7.2 User-centricity

Users will accompany the project from the beginning. User requirements will be collected prior to starting the development, in order to customise the pilot platform and maximise its usefulness. A panel of experts, specialised in grants evaluation process will be put together to accompany the project (e.g. experts from Science Europe). A network of users/stakeholders will be put in place to ensure the future developments stay in line with user requirements and to coordinate issues related to the dissemination and use of the platform.

3.7.8 EXPECTED MAJOR OUTPUTS

3.7.9 ORGANISATIONAL APPROACH

3.7.9.1 Expected stakeholders and their representatives

Stakeholders	Representatives	Involvement in the action
Hungarian Innovation Agency (NKFIH)	Endre Spaller, vice-president	Member of the advisory board, providing expertise in the evaluation process of research proposals, test pilot platform, provide data.
Spanish foundation for science and technology (FECYT) + funding agencies	Cecilia Cabello, director S&T Indicators and R&D and Innovation Policy Monitoring	Member of the advisory board, providing expertise in the evaluation process of research proposals, test pilot platform, provide data.
Science Europe	Stephan Kuster, Acting Director	Member of the advisory board, providing expertise in the evaluation process of research proposals
Joint Research Centre	Olivier Eulaerts, team leader	Member of the advisory board, providing IT expertise (text mining, data, …)
RTD	Common support Centre	Ensure alignment to RTD grant policies + provide data
ERCEA	Alexis Michel Mugabushaka, Head of Sector Monitoring & Evaluation.	Member of the advisory board, providing expertise in the evaluation process of research proposals, test pilot platform, provide data.

3.7.9.2 Identified user groups

Public R&I funding agencies in Member States
Public R&I funding agencies in H2020 Associated States.
R&I agencies at international level.
Applicants to R&I grants.
Private funding agencies.

3.7.9.3 Communication and dissemination plan

Dissemination activities for the pilot phase will focus on informing stakeholders of the existence and objectives of the project. This will be done via the funding agencies themselves and via Science Europe. The group of experts that will accompany the project will be asked to recommend the means of dissemination for a full COMPARED platform, should it go for full deployment. An exhaustive communication and dissemination plan will then be designed, if the pilot phase concludes positively and if the full deployment of the platform is launched. This plan will involve online presence and offline materials, but would probably focus on networking, presentation to dedicated workshops and conferences. Corporate dissemination via the ISA² network of Member States coordinators could also be an efficient means of dissemination.

3.7.9.4 Key Performance indicators

Description of the KPI	Target to achieve	Expected delivery (months after k-o)
Kick-off Workshop	At least 15 experts in evaluation processes for research proposals from public funding agencies from Member States.	+M1

User requirements documents	List of requirements for semantic platform for R&I proposals	+M2
COMPARED pilot platform and testing	Web application accessible + testing by experts from public funding agencies from Member States.	+M11
Closing workshop	At least 15 experts in evaluation processes for research proposals from public funding agencies from public funding agencies from Member States.	+M12
Recommendation report	Report by expert group on full deployment	+M12

3.7.9.5 Governance approach

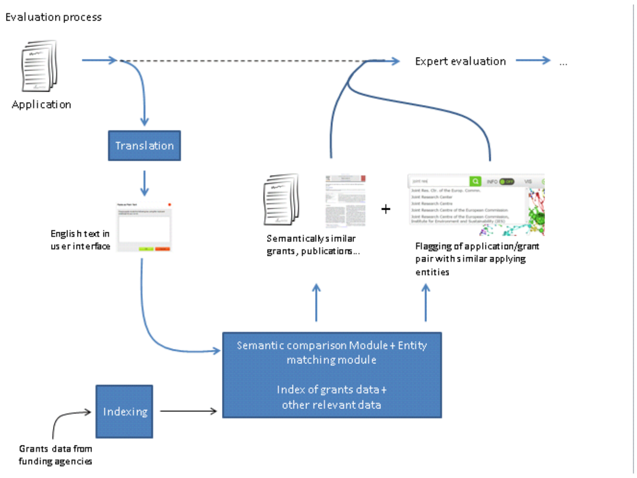
To limit the cost in case of project failure, COMPARED is designed as a two-phase project. At the end of the pilot phase the potentialities, added value and feasibility of scaling up the COMPARED platform will be analysed by a group of experts which will deliver a report containing recommendation for further development and scale-up.

Experts will be involved throughout the whole pilot project: they will set up the specifications for such a system and will evaluate the pilot platform and decide whether it brings sufficient added value for funding agencies to be pursued and scaled-up.

The project will be managed by JRC which will consult and rely on an advisory board composed of representatives from JRC, the Hungarian Innovation Agency (NKFIH), the Spanish foundation for science and technology (FECYT), and Science Europe.

3.7.10 TECHNICAL APPROACH AND CURRENT STATUS

Data used by the COMPARED platform will be indexed (grants, scientific publications, patents). This indexation allows for fast-response checking of incoming proposals against the data. Funding agencies will send their data (or part of it) prior to indexing. The system will be designed for a minimal impact on evaluation processes in Member States: the evaluator will insert the proposal text in an interface that will return a list of matching documents and raise alerts if similar documents are retrieved. Information about applicants will also be provided. In a first instance, proposal texts will be inserted in English. Various solutions for translation will be tested (e.g. MT@EC, Google translation, EMM translation) and offered to the users. The COMPARED platform will be based on text-mining techniques. A first process will measure semantic similarity between proposals for research and a reference dataset, using specific tagging software and cosine distance measurement techniques. A second process running subsequently will identify similar applicants in the submitted proposals and the similar grants that have been retrieved in the first process.


Evaluation process
Application
Expert evaluation
...
Translation
English text in user interface
Semantically similar grants, publications...
+
Flagging of application/grant pair with similar applying entities
Semantic comparison Module + Entity matching module
Index of grants data + other relevant data
Indexing
Grants data from funding agencies

3.7.11 COSTS AND MILESTONES

3.7.11.1 Breakdown of anticipated costs and related milestones

Phase: Initiation Planning Execution Closing/Final evaluation	Description of milestones reached or to be reached	Anticipated Allocations (KEUR)	Budget line	Start date	End date
Initiation and planning	- Kick off workshop - User requirements document	30k€ experts + 32k€ IT	ISA² - JRC	April 2018	May 2018
Execution	- Logistics (meetings, missions) - Platform design, customisation, testing. - Data collection, gathering, formatting, storage, integration, indexing. - Setting up of a network of funding agencies from Member States - Setting up of network of expert evaluators - Interface with funding agencies and business analysis (IT requirements, data requirements, etc.) - Exploration of legal issues related to data access and sharing. - Hardware	339k€ IT +10k€ missions-logistics + 15k€ hardware	ISA² - JRC	April 2018	May 2019
IT supervision	IT supervision and architecture	25k€	JRC	April 2018	May 2019
Closing and Final decision	- Testing of platform. - Closing meeting - Final go / no-go for full deployment.	30k€ experts + 32k€ IT	ISA² - JRC	March 2018	May 2019
	Total	513k€			

3.7.11.2 Breakdown of ISA^2 funding per budget year

Budget Year	Phase	Anticipated allocations (in KEUR)	Executed budget (in KEUR)
2018	Pilot phase	250€	
2019	Pilot phase	160k€	