Proceedings of the First workshop on Language Technologies for Government and Public Admnistration (LT4Gov), pages 22–27
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Research & Innovation Activities' Impact Assessment: The Data4Impact System

Ioanna Grypari, Dimitris Pappas, Natalia Manola, Haris Papageorgiou
Athena Research & Innovation Center
Artemidos 6 & Epidavrou 15125, Marousi, Greece
{igrypari, dpappas, nmanola, haris} @ athenarc.gr

Abstract

We present the Data4Impact (D4I) platform, a novel end-to-end system for evidence-based, timely and accurate monitoring and evaluation of research and innovation (R&I) activities. Using the latest technological advances in Human Language Technology (HLT) and our data-driven methodology, we build a novel set of indicators in order to track funded projects and their impact on science, the economy and the society as a whole, during and after the project life-cycle. We develop our methodology by targeting Health-related EC projects from 2007 to 2019 to produce solutions that meet the needs of stakeholders (mainly policy-makers and research funders). Various D4I text analytics workflows process datasets and their metadata, extract valuable insights and estimate intermediate results and metrics, culminating in a set of robust indicators that the users can interact with through our dashboard, the D4I Monitor (available at monitor.data4impact.eu). Therefore, our approach, which can be generalized to different contexts, is multidimensional (technology, tools, indicators, dashboard) and the resulting system can provide an innovative solution for public administrators in their policy-making needs related to RDI funding allocation.

Keywords: HLT, NLP, RDI, impact evaluation, policy-making, public administration

1. Monitoring, Evaluation & Policy-making in R&I Activities

As the number of programmes available for financing Research & Innovation (R&I) activities has been growing, so has the need to update the monitoring and evaluation of such activities. Traditional assessment systems are costly, rely disproportionately on the self-declared performance from project participants, and limit the relevant evaluation period to the project's lifetime.

Moreover, the accurate assessment of awarded R&I projects is essential for future policy-making. Answers to questions such as:

"What lessons can be learned from past projects? Which research areas are emerging? How much have I, as a funder, invested in those areas compared to other funders, and how "crowded" are they? What "type" of projects create innovations that reach the market quickly? Which companies are still innovating in the same field that they received funding for? Which organisations play a key role in the diffusion of technology? What are the most important research communities and how are they spread out across countries and sectors? What are the characteristics of projects whose outputs reach the average person faster? What issues do people care about?"

among others, are key in understanding the potential impact of R&I activities, allow for evidence-based policy-making and, in principle, for an "optimal" allocation of funding resources.

In fact, there is a wide range of research on the different possible R&I impact avenues and their estimation techniques, the most established of which comes from scientometrics. Nevertheless, technological advancements in the areas of HLT have brought forth the capabilities to update these traditionally-used tools and significantly augment our approach with more varied sources of data and frontier technologies.

There is, thus, an opportunity to build monitoring systems that are, to a large extent, automated and offer accurate, timely, granular and multidimensional estimates of the performance of R&I activities and their effects on the society at large. This is the mission of the Data4Impact (hereafter D4I) platform, which we present in this showcase.

There is limited literature on evidence-based end-to-end systems. STAR metrics (Largent and Lane, 2012)is a US infrastructure that tracks a wide range of administrative and other data to analyze input, output and outcomes of federal R&D investments. Corpus Viewer (Pérez-Fernández et al., 2019), a Spanish initiative, uses HLT technologies on text and metadata (mainly from patents, scientific publications and grant proposals) to build indicators for policy evaluation, and additionally offers tools for policy implementation and identification of cases of double funding and fraud in proposals.

Although *complementary* to the D4I approach, there are several differences among the three systems; the most prominent being the sources and coverage of indicators. In fact, to the best of our knowledge, our platform is the only one that offers a holistic approach and at this level of breadth, supporting indicators from input to the different possible dimensions of impact.

Using HLT and other methodologies, we extract pertinent information from project reports, publications, patents,

company websites, policy documents (clinical guidelines), products (drugs) and traditional and social media and link them across different entities (projects, topics, countries, funders, and so on). This results in a rich database of analytics and indicators that can be "sliced" across different dimensions according to the needs of the policy-makers and other stakeholders.

Moreover, our platform accommodates the D4I Monitor,[1] a BI tool that allows us to map a complex set of methodologies and analytics onto a user-friendly dashboard with interactive visualizations of indicators and customization capabilities.

In Section 2, we briefly describe the datasets, methodology and resulting indicators of the workflows of the D4I end-to-end platform, and proceed, in Section 3, to present the dashboard. In Section 4, we conclude.

2. D4I Processing Workflows & Indicators

Data4Impact is a Horizon 2020 project[2] aimed at addressing the mission described in the Introduction. Namely, we built end-to-end workflows that use the latest technologies in Machine/Deep Learning to create a novel and rich set of indicators that are granular, timely and track a funded project's performance and its impact, well after the end of its life-cycle.

We broke down the monitoring needs of a project into five stages: input (at the initial setup of the project), throughput/output (during/at the end of the project) and academic, economic and societal impact (mostly after the end of the project capturing its mid- and long-term impact). We developed our methodology by focusing on EC projects in health and health-related fields in FP7 and H2020 programmes. Importantly, experts in the particular sector guided us to the right data sources for identifying the input-to-impact story. Our approach is generalizable to other policy areas, conditioned upon the human-in-the-loop process to guarantee good coverage of impact scenarios.

2.1. Projects

In order to track the input-to-impact process across projects, we start by examining the textual content of project-related documents (associated call, proposal, reports, deliverables, publications and patents created in the context of the project, and so on). We use a wealth of NLP methods in the steps of our workflows.

First, we segment the content of project reports and publication abstracts (i.e. using its *rhetorical structure*) and isolate the sections and publication zones that relate to the contributions, results and impact of each project and research team. Next, we conduct entity recognition and keyterm extraction using SGRank (Danesh et al., 2015) to help define the work conducted and subject matter of each document, and to build graphs that depict the correlations between entities. This allows us to explore spatial/temporal trends and patterns across projects. Further, we apply our innovation extraction framework that annotates innovation statements into a pre-defined set of domain-independent (e.g., publication, patent, employment), domain-related (e.g., device, diagnostic tool) and domain-dependent (e.g., drug, treatment, clinical trial) insights.

Moreover, it is important to note that disease mentions along with MeSH terms and other established disease classification schemes, like the ICD, are leveraged to automatically classify projects according to research areas. Additionally, and with the objective to provide multidimensional KPI analytics, metadata are taken into account. Specifically, financial data about the cost of each project along with the budget distribution per participant are considered. Moreover, data relevant to each organisation participating in the project, such as the country it is based and its type, i.e., whether it is a research organisation, a university or a company, is gathered and leveraged to construct collaboration networks that help quantify the collaboration and diffusion of technology (using different centrality measures) between the beneficiaries. Merging this work with the extracted data analytics and classification enable us to create a wealth of indicators that can be compared across different types of entities such as funders, time, participating organisations, etc.

To track the evolution of innovations and measure the impact of projects past their life-cycle, we target different data sources. First, we measure the technological value of patents produced in projects by counting their forward and backward citations,[3] and the technological value of publications by examining how many patents cite them.

Second, to build economic impact indicators we crawl the websites of the private-for-profit beneficiaries in the projects. We apply our NLP workflows and pipelines as described above, adapted to the task, isolate their current innovation activities and outputs, and quantitatively relate them to those produced in the context of their EC projects. In particular, we propose a novel method to proxy the commercialization of projects' innovations by the companies, the *uptake score*, by creating a graph semantically linking the keyterms from the three types of documents (project reports, publications, company websites).

Third, to examine the societal impact of projects, we collect policy documents (clinical guidelines), clinical trials and data related to drugs linked to projects. We analyse the contextual fragments related to cited references and other extracted data to construct indicators that measure the reach of the project innovations to the society via generating health-related impact.

[1] Under development and available at monitor.data4impact.eu.
[2] cordis.europa.eu/project/id/770531
[3] I.e., the number of patents a particular patents cites, vs. the number of patents that cite the particular patent

2.2. Topics

Understanding the need of policy-makers to assess the input-to-impact process also from a "bird's eye" view, we worked on the training and development of topic models. The Multi-View Topic Modelling framework consists of several components targeting information extraction, semantic annotation and, most importantly, automated multi-dimensional analysis based on an innovative multi-view probabilistic topic modelling engine (Metaxas and Ioannidis, 2017). We took a large sample of health-related research and used this bottom up approach to divide the field into topics that were manually validated and labeled by a field expert, and placed into generic, major categories. The output of the topic modelling algorithm also provides us with project-topic associations. This allows us to connect all the project-level data, and the previously-described indicators, with their topic distribution. It is also the key in the construction of "non-traditional" academic impact indicators that measure the timeliness, investment potential and exclusivity of research funding, amongst other variables, by comparing the strength of a topic (research volume) across different funders and the entire (academic) health domain. Further, the richness of the topic modelling output, together with the metadata available, allow us to create topic-based similarity indicators that enable us to compare different entities (e.g., countries) according to the topic distribution of their research output.

Lastly, to expand our analysis of societal-level indicators, we pick a subset of "essential" topics (determined using project extracted innovations and the insight of a field expert), and performed relevant searches on traditional and social media.

In particular, we gauge the societal relevance of these topics, by creating indicators based on their media buzz as well as on different characteristics of twitter conversations related to them. The latter is also augmented with visualizations that depict the evolution of twitter discussions. (Lorentzen et al., 2019).

2.3. The Input-to-Impact Story

This rich set of metrics and indicators is thus supported by establishing links among different types of entities at a granular level. First, by aggregating hierarchically upwards we can examine R&I activities and their impact at the Project, Call, Programme and Funder (e.g., the EC) level.[4] Second, using the project metadata we can refine the results further and filter them for particular organizations (beneficiaries), countries and over time.

Moreover, we are also able to track R&I activities from input to impact at the very fine level of the topic (or aggregated to a major category or the entire health field). This offers a different view of monitoring that is well-suited for comparisons across different entities as it abstracts from the programmatic structure of funding schemes, and can also offer a rich and novel set of indicators that rely on topics and their characteristics. Further, through the project-topic associations, topic-based indicators can be also refined and examined for particular organizations (other funders or project participants), countries and over time.

Therefore, one of the strengths of the D4I indicators, and the underlying workflows, lies on the fact that the input-to-impact story of R&I activities can be unfolded in two ways: via projects or via topics (and the entities hierarchically above each).

Lastly, these novel indicators, in combination with traditionally used ones, can provide policy-makers with the quantitative information needed to conduct an in-depth and well-rounded assessment of various investment/funding opportunities by examining the correlations of metrics and project characteristics across different project stages. In other words, these indicators can be used in a statistical analysis not only to answer such questions as the ones presented in Section 1, but also to formally show the interplay among them.[5]

3. D4I Monitor

Given the complexity of the processing workflows and the various data sources, and in order to maximize the reach of the newly developed indicators, it is essential to create a flexible and user-friendly dashboard in order to communicate the results to policy-makers. This is the starting point of the D4I Monitor.

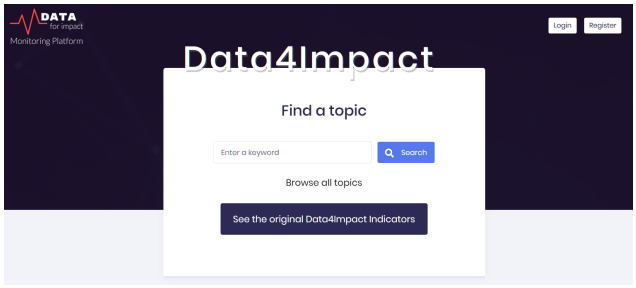

Figure 1: D4I Monitor - Landing Page

The D4I Monitor is an end-to-end Business Intelligence data and visualization tool that can be integrated to third-party platforms. It consolidates the outputs of the different modules of the D4I platform and allows policy operators to interact with and download visualizations and indicators for each of the five stages of input-to-impact described above.

[4] This is the particular hierarchy followed by the EC projects; in general, our approach is adjustable to other funding structures.

[5] As a simple example, one can examine if funding research on emerging topics could also mean contributing in the creation of innovations that not only reach the market quickly but also are successful, in the sense of people knowing and using them.

We organise the data on the dashboard to fit the needs of our stakeholders. In particular, a user can view a report, i.e., a series of visualizations displayed over five input-to-impact tabs, by first selecting either a topic (or major category, or field), or an item from their portfolio (Search Bar in Figure 2). The latter is populated with projects, calls, programmes and organisations that the user selects, conditioned on data access rights (Figure 3).[6]

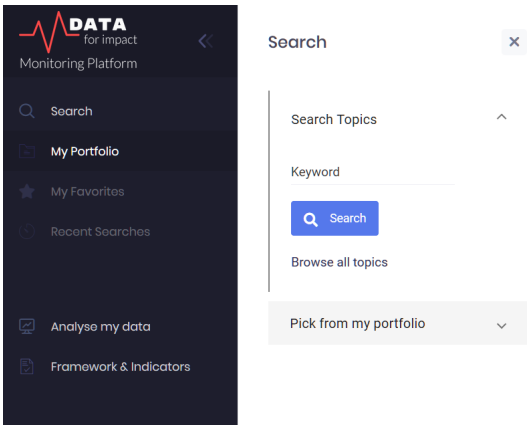

Figure 2: D4I Monitor - Side Bar & Search Bar

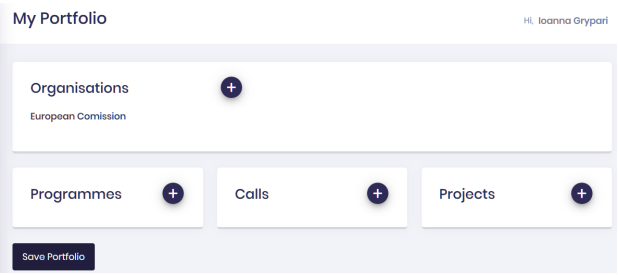

Figure 3: D4I Monitor - Portfolio

In order to use all pertinent data we have available, at the topic search, we match the word typed by a user not only to the name of a topic, but also to the associated keywords/phrases from the topic models and rank topics by quality of match using the corresponding keyword weights. As an example, Figure 4 displays the results that come up after searching for the keyword "malaria."

Once a particular entity is selected and the report is displayed, the user can take advantage of our multidimensional analysis by filtering the entire report further by the country, participating organisation or time range of interest (Figure 5).

In the report itself, each interactive visualization presents the values of one or more indicators. A user can filter

[6]Given a different funding structure (e.g., personal grants), the portfolio would be adjusted to list the corresponding funding levels.

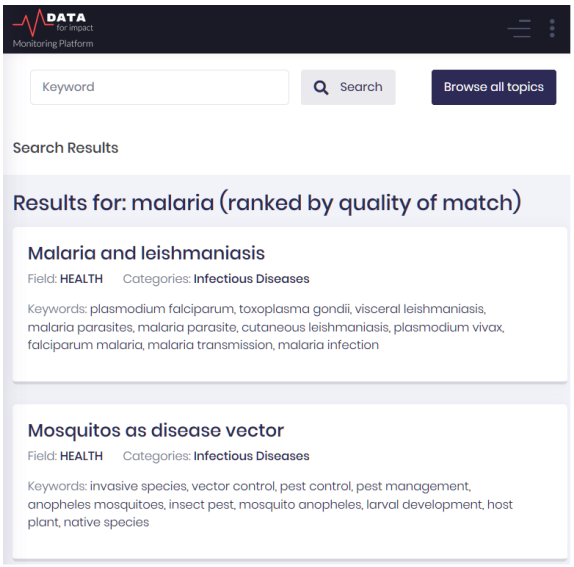

Figure 4: D4I Monitor - "malaria" Search Results

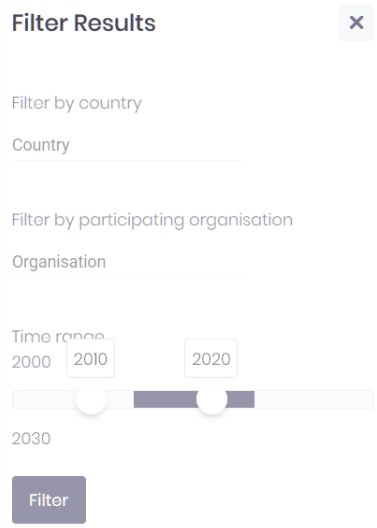

Figure 5: D4I Monitor - Filtering (Not Active)

for particular entities and hover to view values of interest. There is the option to download the entire report in PDF and each visualization separately in PNG (the filtered/zoomed-in image), and the data behind it in CSV or JSON file formats (Figure 6).

Further features are being built so that the dashboard can meet the requirements of a go-to monitoring tool for R&I activities. In particular, users will be able to save and monitor different entities, receive updates, and request to have their own data analyzed and the results uploaded on the platform (Side Bar in Figure 2). In addition, the D4I Monitor is flexibly built so that it can accommodate more fields and indicators.

There are currently pilot studies underway, most recently with policy-makers working on rare diseases, to continue

the improvement of the dashboard, the indicators and the underlying technologies.

4. Conclusion

In summary, the D4I platform brings together the information from a variety of sources and applies state-of-the-art methods to derive meaningful, timely and reproducible indicators linked across different entities. The developed end-to-end system allows stakeholders to monitor and evaluate their funding schemes and conduct data-driven policy-making. Our end-product, the D4I Monitor, is a user-friendly and agile platform that warrants ease of access of the results to policy-makers and guarantees the continued improvement of their policies.

5. Acknowledgements

Data4Impact has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770531. The authors would like to thank all the consortium partners of Data4Impact for a fruitful and rewarding collaboration, namely PPMI, Fraunhofer ISI, CNR, University of Borås and Qualia SA.

6. Bibliographical References

Danesh, S., Sumner, T., and Martin, J. H. (2015). SGRank: Combining Statistical and Graphical Methods to Improve the State of the Art in Unsupervised Keyphrase Extraction. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, SEM 2015*, pages 117–126.

Eklund, J., Gunnarsson Lorenzen, D., and Nelhans, G. (2019). Mesh Classification of Clinical Guidelines Using Conceptual Embeddings of References. In *17th International International Conference on Scientometrics and Informetrics, ISSI 2019*, volume 2, pages 859 — 864.

Feidenheimer, A. and Stanciauskas, V. (2019). Developing KPI's for Impact for Institutions and (Inter)national Policies. In *Impact of Science 2019*. Presentation.

Fergadis, A., Baziotis, C., Pappas, D., Papageorgiou, H., and Potamianos, A. (2018). Hierarchical Bi-directional Attention-based RNNs for Supporting Document Classification on Protein–protein Interactions Affected by Genetic Mutations. *Database*, 2018.

Grypari, I., Nelhans, G., and Stanciauskas, V. (2019). Application of Big Data in Scientometrics. In *17th International Conference on Scientometrics and Informetrics, ISSI 2019*. Workshop.

Gurell, J. and Nelhans, G. (2018). A National CRIS in Sweden – a Developer's and a Researcher's Perspective. In *CRIS2018: 14th International Conference on Current Research Information Systems*. Keynote Address.

Largent, M. A. and Lane, J. I. (2012). STAR METRICS and the Science of Science Policy. *Review of Policy Research*, 29(3):431–438.

Lorentzen, D., Eklund, J., Nelhans, G., and Ekström, B. (2019). On the Potential for Detecting Scientific Issues and Controversies on Twitter: A Method for Investigation Conversations Mentioning Research. In *17th International Conference on Scientometrics and Informetrics, ISSI 2019*, pages 2189 – 2198.

Metaxas, O. and Ioannidis, Y. (2017). Multi-View Topic Modelling on Text-Augmented Heterogeneous Information Networks. Under submission.

Nelhans, G., Vlachos, E., and Vigne, M. (2019). Towards a Responsible Institute Impact Assessment. In *Liber Conference 2019*. Presentation.

Pérez-Fernández, D., Arenas-García, J., Samy, D., Padilla-Soler, A., and Gómez-Verdejo, V. (2019). Corpus Viewer: NLP and ML-based Platform for Public Policy Making and Implementation. *Procesamiento del Lenguaje Natural*, 63:193 – 196.

Pukelis, L. and Stanciaukas, V. (2018). Big Data Approaches to Estimating the Impact of EU Research Funding on Innovation Development. In *STI 2018 Conference Proceedings*, pages 429 – 435.

Pukelis, L. and Stanciauskas, V. (2019). Opportunities and Limitations of Using Artificial Neural Networks in Social Science Research. *Politologija*, 94(2):356 – 380.

Pukelis, L. (2019). Using Internet Data to Compliment Traditional Innovation Indicators. In *International Conference on Public Policy (ICPP4)*. Presentation.

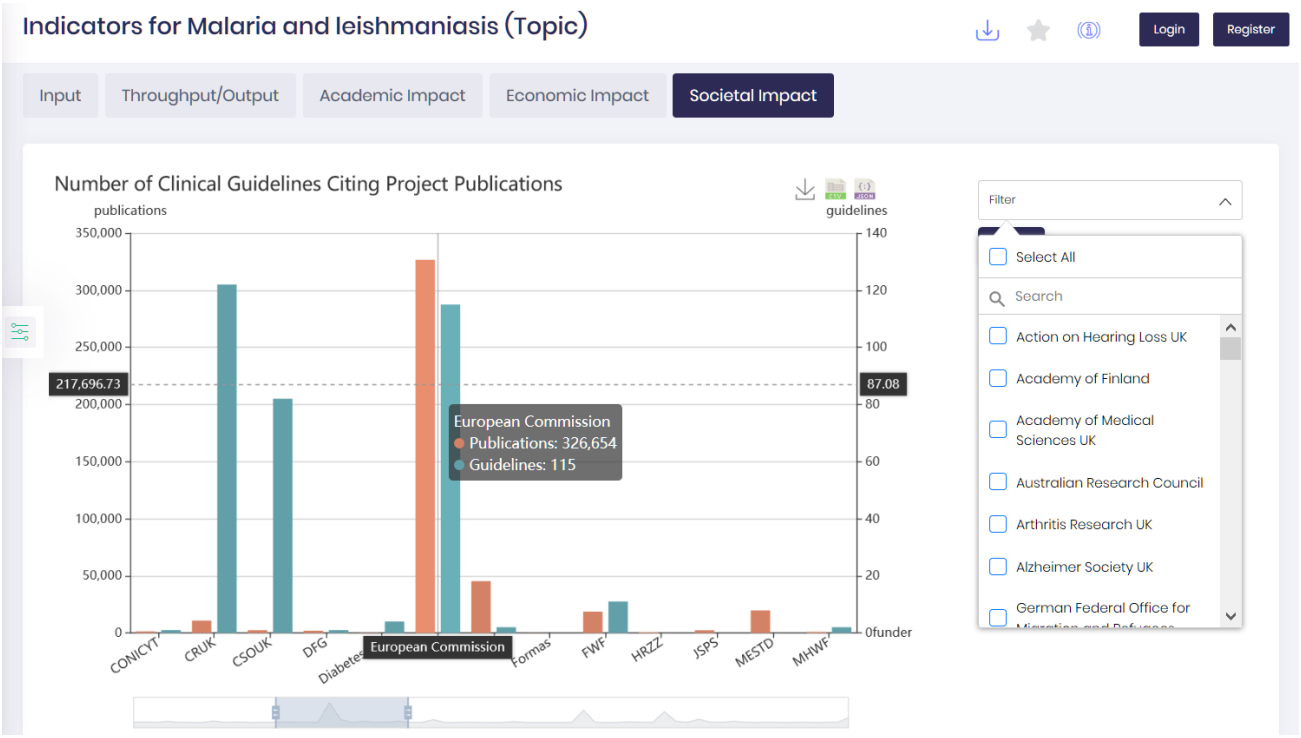

Figure 6: D4I Monitor - Sample from a Report