# Workshop on Research Policy Monitoring in the Era of Open Science and Big Data

## The what (indicators) and the how (infrastructures)

27-28 May 2019

Ghent, Belgium

## Workshop report

OpenAIRE-Advance receives funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 777541. Data4Impact receives funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 770531.

# 1. Introduction

Research funders across Europe are increasingly mandating Open Science practices for funded research outputs to support open and free access to valuable elements of the scholarly communication life-cycle. From OA to publications and the recent PlanS developments, to the promotion and uptake of coordinated RDM practices, to the more advanced research assessment exercises to understand innovation and societal impact, there is a need for monitoring of research output.

National and EU e-Infrastructures respond to these needs by embedding and developing monitoring tools to provide evidence-based data on policy uptake, costs, and research impact, while at the same time promoting interoperability of information outputs, shareable across networks.

This two-day workshop, co-organised by OpenAIRE and Data4Impact, with support of Science Europe, explored mechanisms for research policy monitoring and indicators, and how to link these to infrastructure and services. The first day was focused on open science indicators as these emerge from national and EU initiatives, while the second day explored more advanced aspects of indicators for innovation and societal impact. Transparency, talking to one another and quality were just some of the key themes which emerged over the course of the workshop, which pooled over 90 research funders, policymakers, experts, infrastructure providers and other members of the Open TDM community.

Other elements addressed in this two-day workshop were:

- Existing ways of monitoring for Open Science – the what and the how
- Collaboration aspects to achieve a seamless monitoring landscape via open infrastructures
- Data driven techniques for research assessment and their links to open data

# 2. Day 1: Monitoring and Infrastructure for Open Science

The first **OpenAIRE-led workshop day** focused on the different ways of monitoring in this new landscape of open science. OpenAIRE, which is a Horizon 2020 project funded by the European Commission, has been working in the area of monitoring and tracking research output, especially that of funders for nearly ten years now.

## 2.1. Summary of Day 1 morning session

### 2.1.1. Quality must come first

The forward-looking keynote speech, given by Marc Vanholsbeeck, explored the meaning of impact. What we are referring to by using the term? There are, in fact, so many kinds of impact in research, over 3000+ pathways according to the available research. Therefore, it is hard to pin down what we want to measure. The overarching theme was: quality comes first, impact comes next. It goes without saying that policies, such as those coming from the EC or national funders can determine impact, but how do we measure it?

What is certainly clear is that OS can both communicate and popularize research to the benefit of society – as well as drive forward new ways of measuring impact. The remaining question is who are the players and providers who can do this – as an offer to the society?

For more information on this presentation, click **here**.

### 2.1.2. Need to monitor repository and get training

A presentation of the OS monitor furthered the discussion – research by the initiative has looked into incentives for researchers to share. Very few researchers are willing to share their research data beyond

their research groups. One interesting point made was the need to standardize usage data coming from repositories. A clear way forward is to train data stewards, the number of whom is still very few.

For more information on this presentation, click **here**.

#### 2.1.3. Standartisation of FAIR

The RDA-FAIR Data Maturity Model WG - outlined in a presentation by Brecht Wyns and Christophe Bahim - has set out a rich plan looking at how we standardise FAIR - at present there are no benchmarks. This WG will go far to set these criteria and interpret the implementation of FAIR. The landscape study will also be crucial. We look forward to that FAIR checklist.

For more information on this presentation, click **here.**

#### 2.1.4. CoalitionS – part of a global movement

The Coalition S members have had their hands busy taking on board all the many responses. As a result, there will be some changes altering the existing key principles. This includes: no pay-walls, more flexibility on the CC-BY licensing and a bit of leeway on allowing hybrid, so long as the journal can demonstrate it is moving towards a full open access model. And…the green road is also a very important route to open access reflecting that CoalitionS is part of a global movement. Good to see! There are few sanctions, the ultimate goal, here, is to change the publishing system, not to punish the researchers.

Important point from Science Europe – remember that not all funders are so well resourced to deal with all of these monitoring issues.

## 2.2. Summary of Day 1 afternoon session

#### 2.2.1. Indicators should be transparent.

Dietmar Lampert's presentation stressed that - above all - indicators should be transparent. The results presented from his study (ZSI Research Policy and Development) gave some interesting insights: what indicators should be developed such as:

% of Pubs of OA journals

- % of Pubs of OA journals with no impact factor
- Availability of means to easily publish negative results

Researchers are all too aware of standards. Therefore, it is natural that we can build one for OS. Other factors can give us monitoring insights, therefore we need to explore them. Research evaluation: the unseized opportunities of the open science era and the possible technology and methodology approaches.

For more information on this presentation, click **here**.

#### 2.2.2. Measuring Open Science

The presentation of Diego-Valerio Chialva (European Research Council) examined the potential of the semantic web and using linked data to measure OS, such as locating a resource, and its relationships.

For more information on this presentation, click **here**.

#### 2.2.3. OPERA: Exploring Open Research Analytics

This could also be the case highlighted in the OPERA project (Karen Hytteballe Ibanez, Technical University of Denmark). The OPERA project: Exploring Open Research Analytics using VIVO. In this project, Vivo can be used to describe open science metrics along with collaborations international and national, funding national and international, importantly using VOSviewer, and NEO4J for visualisations.

Ultimately, they are dependent on open sources and transparency. They are reliant on a number of sources.

For more information on this presentation, click **here**.

#### 2.2.4. It is all about good data

Paolo Manghi's presentation also reiterated the need for reliable data to contribute to the OpenAIRE graph. Much of this graph looks beyond the article itself such as: funder data – projects – research data – software which leads to a huge potential of added value services. And it needs to be kept curated and clean. Ultimately, we are building this graph together as decentralised open science community from repositories, aggregators, individuals, institutions, service providers. Ultimately all can contribute, and this would be for the common good. However, the data needs to be reliable.

For more information on the presentation, click **here**.

#### 2.2.5. An inspiring tale from Portugal

RCAPP started in 2008 which provides access to all HEI material nation-wide. Involving pretty much everyone from the scholarly communication chain, it serves up a number of services. Using OpenAIRE's services, project information is also automatically added to output. It works as a successful case study whereby deposit is a natural part of the lifecycle within the institutional setting and since standards are implemented early on, the researcher only has to deposit once and reap the benefits.

For more information on this presentation, click **here**.

## 2.3. Summary of Day 1 panel discussion

The workshop programme concluded with a **panel discussion** involving the following representatives of the organisations which serve as infrastructure providers, funders and policymakers:

- Bregt Saenen (EUA)
- Helena Cousijn (FREYA)
- Paolo Manghi (OpenAIRE)
- James Hardcastle (Clarivate Analytics)
- Diego- Valerio Chialva (ERC)

The panel discussion was moderated by Natalia Manola, OpenAIRE Managing Director, ATHENA Research & Innovation Center.

The panel discussion focused on exploring the following key questions: How can we work together? Also, who pays: this - in itself - is a negative question. It should be framed as 'what is the cost going to be in comparison to the existing system'. It was discussed that it is important to be interoperable to work together. The systems and their outputs have to interact because interaction will make thing easier. Two other panel members stressed that if we find similar ways to organise and discuss ideas, then this sort of discussion can continue.

Some panel members argued for a consortia approach. The single institution cannot pay alone. Against the backdrop of the audience mentimeter, it was clear that 'openness has to somehow be monitored. However, it was clear that lack of trusted sources is the main barriers. It was a start.

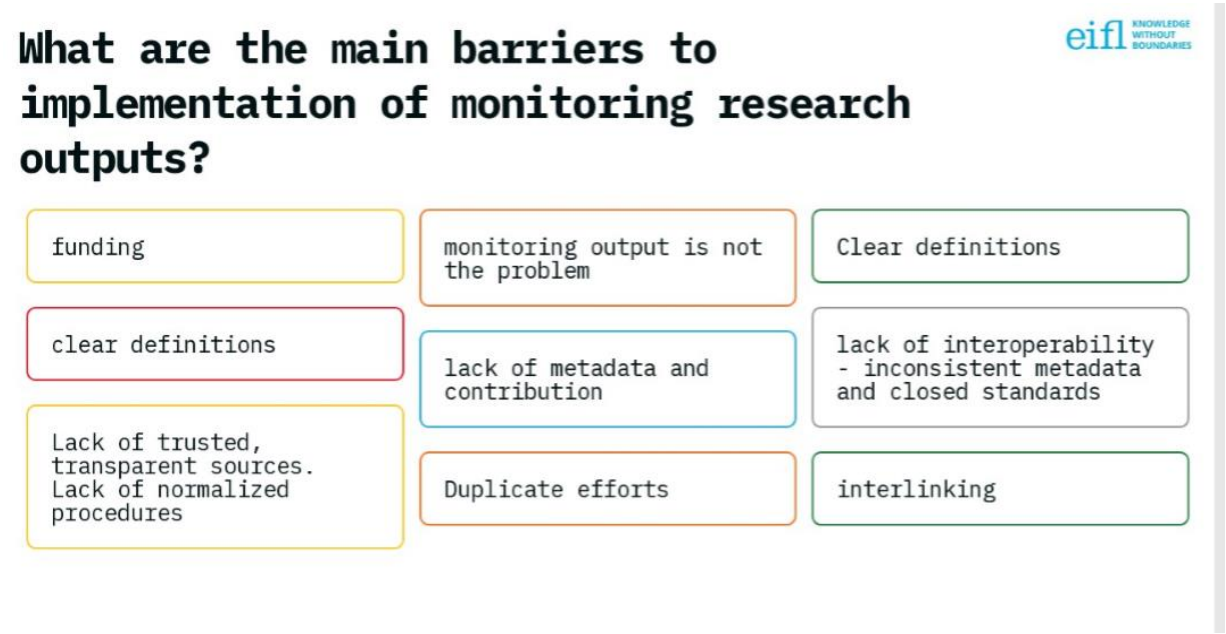


**Figure 1. Main barriers to the implementation of research outputs monitoring**


**Figure 2. Key aspects of research which need to be monitored**

Some of the main conclusions of the panel discussion were:

- Do not collect numbers just for the sake of it.
- We need to carefully construct the parameters. This can lead to perverse results.
- We need a quality trust seal for research outcomes and data.
- We need benchmarking for open science.
- It has to be transparent.
- We need to be able to easily compare policies.

# 3. Day 2: Open Science and Big data in support of measuring R&I Indicators

The second **Data4Impact-led workshop day** focused on the use of big data technologies for advanced research assessment. Data4Impact, which is a Horizon 2020 project funded by the European Commission, pioneers big data techniques and develop pilot approaches tracking legacy and impact of research activities after the end of public funding. In this workshop, the project consortium presented a series of indicators developed on the performance and societal impact of 40+ research programmes in the health domain for the first time.

## 3.1. Summary of Day 2 morning session

### 3.1.1. Introduction of Data4Impact

Data4Impact coordinator Vilius Stanciauskas (PPMI) introduced the project. The key message was that building on data harvested from PubMed, OpenAIRE, Lens.org, PATSTAT, clinical guidelines repositories, company websites, social media and media platforms, EC monitoring data and other databases, Data4Impact enables policymakers, funders, experts, researchers and the public in general to **'Ask less & know more'** in the context of advanced research assessment.

### 3.1.2. Data4Impact Analytical Model Of Societal Impact Assessment (AMOSIA)

Following that, Alexander Feidenheimer (Fraunhofer ISI) provided an overview of the Data4Impact Analytical Model Of Societal Impact Assessment (AMOSIA) which has been developed over the course of the project. The analytical model is structured around four distinct phases of the research lifecycle, including **input, throughput, output, and impact**. Relying on novel big data techniques such as web scraping, crawling and mining as well as text analysis methods such as Natural Language Processing and deep learning, Data4Impact has gathered data for each analytical phase.

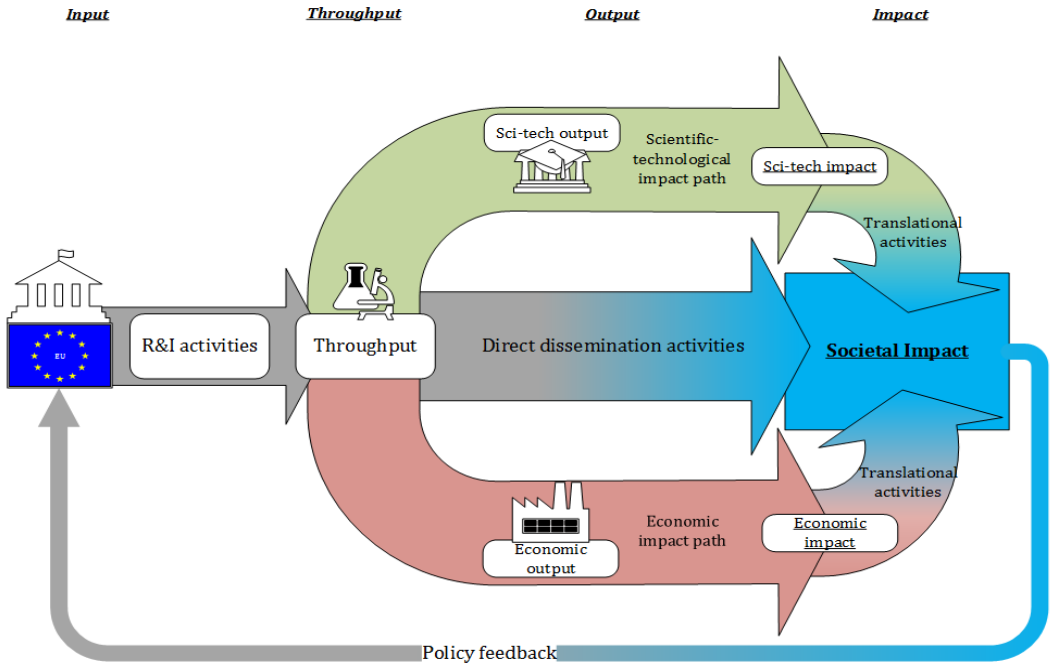


**Figure 3. Data4Impact Analytical Model Of Societal Impact Assessment**

### 3.1.3. Tracking of Research Outputs

The primary purpose of the following presentations by Data4Impact consortium was to present the indicators which have been developed over the course of the project for each of these phases. Ioanna Grypari (ATHENA RC) discussed the approach used for **tracking of research outputs**.

### 3.1.4. Academic, Economic & Societal Impact

Then, the consortium including Ioanna Grypari (ATHENA RC), Iason Demiros (Qualia), Vilius Stanciauskas (PPMI) and Gustaf Nelhans (University of Borås) have put emphasis on three categories of impact, including **academic, economic and societal impact**. Please refer to the **Data4Impact booklet** which demonstrates linkages between data collected across the three dimensions of impact **here.**

For more information on the second workshop day presentation, click **here.**

## 3.2. Summary of Day 2 afternoon session

### 3.2.1. Group discussions

Data4Impact invited the participants to two parallel sessions, focusing on the Data4Impact methodology and indicators in the areas of:

- Academic Impact & Societal Relevance of Research
- Economic Impact & Societal & Health Impact

The following table summarises the key Data4Impact indicators which were presented to the discussion groups as well as their assessment of the utility and credibility of indicators in each of the focus areas.

**Table 1. Summary of Data4Impact indicators & their assessment by the audience**

| Level | Indicator | Description | Relevance & Credibility | Comments |
|---|---|---|---|---|
| Academic impact | Funding priorities | Topic size in PubMed (absolute and normalised) | High | N/A |
| | | Distribution of topics per funder (normalised) | High | N/A |
| | | Distribution of funders per topic (absolute) | High | N/A |
| | Timeliness of research performed | Rate of topic growth between 2012-2018 compared to 2005-2011 | Low | The length of the time intervals used to estimate trends, growth and other factors would be derived by some well-established and/or relevant criterion |
| | | Share of funding allocated to top-10% fastest growing topics | High | N/A |
| | Funding exclusivity | Share of funding allocated to top-10% smallest research topics by size (i.e. investment in small-niche topics) | Medium relevance but high credibility | N/A |
| | | Number of funders per topic whose output share exceeds 3% globally | Medium | To be clarified |
| | | Share of funding allocated to research topics with less than 5 funders whose share exceeds | Medium | To be clarified |

| | | | | |
|---|---|---|---|---|
| | | the 3% mark (i.e. investment in topics where few other funders invest) | | |
| | Technological value/significance of patents | Analysis of the extent to which commonly patent forward citations (i.e. citations a patent receives from subsequent patent filings) are used | High | N/A |
| Economic impact | Economic and innovation performance of companies | Estimated share of enterprises with evidence of innovation activities | High | Might fit better if presented as input level indicator |
| | | Estimated share of highly innovative enterprises | High | |
| | | Estimated share of enterprises with evidence of licensing activities (incl. patent/ trademark license agreements) | High | |
| | | Estimated share of enterprises involved in activities related to acquisitions | High | |
| | | Estimated share of enterprises with evidence of private investment/capital attracted | High | |
| | Continuity of innovation activities | Estimated overlap between project activities in FP7 & identified company innovations | Medium | To be clarified |
| | | Number of newly CE-marked devices and medical technologies that could be directly linked to R&I activities in the EU Framework Programmes | Medium | |
| Societal/ health impact | Impact on public health | Citations of publications in clinical guidelines | Medium | Might be relevant to consider policy impact (e.g. consider health technology assessment) |
| | Societal awareness/relevance of research | Rank of research topic based on number of news articles, blogs, posts, tweets, etc. discussing a given topic | Low | N/A |
| | Congruence of research funding with societal priorities | Rank similarity of most discussed research topics versus actual spending in the topics | Low | N/A |
| | Newly launched medicines and medicinal products | Number of human medicinal products or orphan medicines that could be directly linked to R&I activities in the EU Framework Programmes | High | Most positively assessed indicator |
| | | Strength of link based on the number of mentions of product names and their active substances in EC monitoring data | High | N/A |

Following the discussions on the indicators developed by Data4Impact, the discussion leads from the consortium invited the discussion groups to consider whether any other indicators could be relevant in the context of research assessment. The Academic Impact & Societal Relevance of Research discussion group suggested to explore the possibility to compare the topic of the call with the topics of the publications which came out of the project. This could help determine whether the research published matches the research promised in terms of topics. It was also suggested to consider an indicator which would normalise the numbers produced to enable cross-disciplinary comparisons.

## 3.3. Summary of Day 2 panel discussion

The workshop programme concluded with a **panel discussion** involving the representatives of the funding organisations analyses in Data4Impact as well as service providers including:

- Amit Prasad (World Health Organisation)
- Danil Mikhailov (Wellcome Trust)

- Silvia Pozzi (Fondazione Telethon)
- Frank Manista (JISC)
- Natalia Manola (OpenAIRE)

The panel discussion was moderated by Vilius Stanciauskas, Director of Research & Policy Advice at PPMI and the coordinator of Data4Impact.

The panel discussion focused on the following themes:

- Possibilities and limitations of big data
- Lessons learned from the use case
- Transfer of knowledge and methodology to more environments and programmes

Some of the key conclusions of the panel discussion were:

- Big data shows huge long-term potential, although limitations must be considered (e.g. time lag, data availability, etc.).
- Some ways to mitigate the limitations could be with the use of multiple approaches (e.g. quantitative and qualitative) and additional documentation on how the indicators are derived and what specific data is used to construct them. With increased transparency, policymakers and funders could reach a consensus on the definition of each indicator as well as potential solutions to key data issues.
- Reproducibility of the results considering particularly the usability and ease of use which are the driving force behind the adoption of new technologies.
- Key lessons learnt is that Data4Impact shows that we now have a way to discover and use the data that is out there but could not be collected and used before for research assessment.
- Consider the impact of the fact that some information cannot be accessed openly on the project results.
- Next step is to extend the Data4Impact knowledge and methodology to other domains and programmes.

If you are interested in learning more about Data4Impact methodology and results, we would like to invite you to attend our upcoming **Workshop on Data4Impact Methodology and Indicators** which will take place on **24 June 2019 at the premises of Research Executive Agency** (Brussels, Belgium). In this hands-on, interactive workshop we aim to gather feedback on the chosen methodology, coverage and latency/timeliness of the developed indicators, to maximise their relevance for all the stakeholders involved. You will find more details on the workshop in the PDF file attached. You may find the programme and registration page for this workshop our website. If you have any questions about the event, please contact Sonata Brokeviciute at sonata@ppmi.lt.