Interim DI4P platform operation report

Big Data approaches for improved monitoring of research and innovation performance and assessment of the societal impact in the Health, Demographic Change and Wellbeing Societal Challenge

Project acronym: Data4Impact

Grant Agreement no: 770531

Deliverable: D3.2

Deliverable information	
Deliverable number and name	D3.2 Platform high-level architecture and data flows
Due date	Month 12
Delivery	Month 12
Work Package	WP3
Lead Partner for deliverable	CNR
Author	Paolo Manghi (CNR), Claudio Atzori (CNR)
Reviewers	Omiros Metaxas (ATHENA RIC) and Gustaf Nelhans (UoB)
Approved by	All partners

Dissemination level	Confidential
Version	1.0

Document revision history

Issue Date	Version	Comments
22/10/2018	0.1	First draft.
31/10/2018	1.0	Submitted version.

Disclaimer

This document contains description of the **Data4Impact** project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium coordinator for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of Data4Impact consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (http://europa.eu.int/)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770531.

Abbreviations

DoW – Description of Work
EC – European Commission
PPMI – Public Policy and Management Institute
UoB – University of Borås
ARC – Athena Research and Innovation Center
WP – Work Package
D4IP - Data4Impact Knowledge Base Platform

Table of Contents

Summary

This document is a follow up of the architectural document D3.1 and describes how the platform will be effectively used by the D4I partners to integration of all the information (raw, intermediate, indicators) needed to allow the platform to support: data access capabilities (visualization and bulk download of the indicators, in synergy with WP6) and facilitate the data integration tasks among WP4 and WP5.

CNR prepared this document as the partner leading WP3 with inputs from other consortium partners and with a review from ATHENA RIC and UoB.

1. Introduction

In this document we present the first release of the D4I Knowledge Base Platform (D4IP), whose design has been described in deliverable D3.1. The D4IP is a system capable of supporting the automatic computation and integration of indicators about the impact of science, technology and innovation policies. The platform offered a number of integration patterns for data analysis workflows, whose implementation implied a trade-off between feasibility/cost of the software integration (i.e. external systems, services, and software migrated into the D4IP) and ability to automatically orchestrate all workflows to produce/refresh indicators. This deliverable will present, in particular:

1. The data analysis workflow integration patterns supported chosen by the partners in integrating their analysis workflows;
2. The integration data model devised in order to support partners at implementing their workflows and at integration the results of their analysis, to produce uniform and correlated impact indicators.

As stated in D3.1, the resulting D4IP will operate as the common deposition and access source for input and output data of data analysis workflows and the source of aggregated data interrogation and calculation of indicators. Three main kind of data sources will be integrated to serve as “raw” and intermediate data for data analysis workflows:

- Funder and grant data: programmes, projects; national and EU level;
- Scientific output: article metadata and full texts, patents metadata and full texts;
- Impact data: company data; monitoring data on finalised EU projects; online/social media blog and health-related forum data; and policy documents / data.

Once calculated, the data resulting from the analysis workflows will be made available for programmatic access to end user interfaces to be developed in the context of WP6.

2. Integration of data analysis workflows

As shown in Figure 1, the D4IP allows for the following integration patterns:

Data source integration: harvesting common data sources and output data sources of a local processing workflow.

- Common data sources: OpenAIRE projects and PubMed. In this case the data source is directly harvested by D4IP and offered to the task leader’s local infrastructure to download or to D4IP workflows to reuse.
- Output data sources: the data is collected and processed by local infrastructures in the context of WP4 and WP5 workflows. The solution has low technological integration cost, but does not easily support “continuous generation of indicators”. Agreements on the frequency of harvesting must be taken between WP output data production and D4IP harvesting workflows.

Workflow integration: integration of business logic into D4IP, requires technology integration into the platform workflow engine. The platform will then harvest the primary data and execute

the process locally. This solution supports “continuous generation of indicators” but at the cost of integrating heterogeneous technologies.

External workflow integration: external workflows can be executed by the D4IP. This solution supports “continuous generation of indicators” mitigating the cost if integrating heterogeneous technologies.

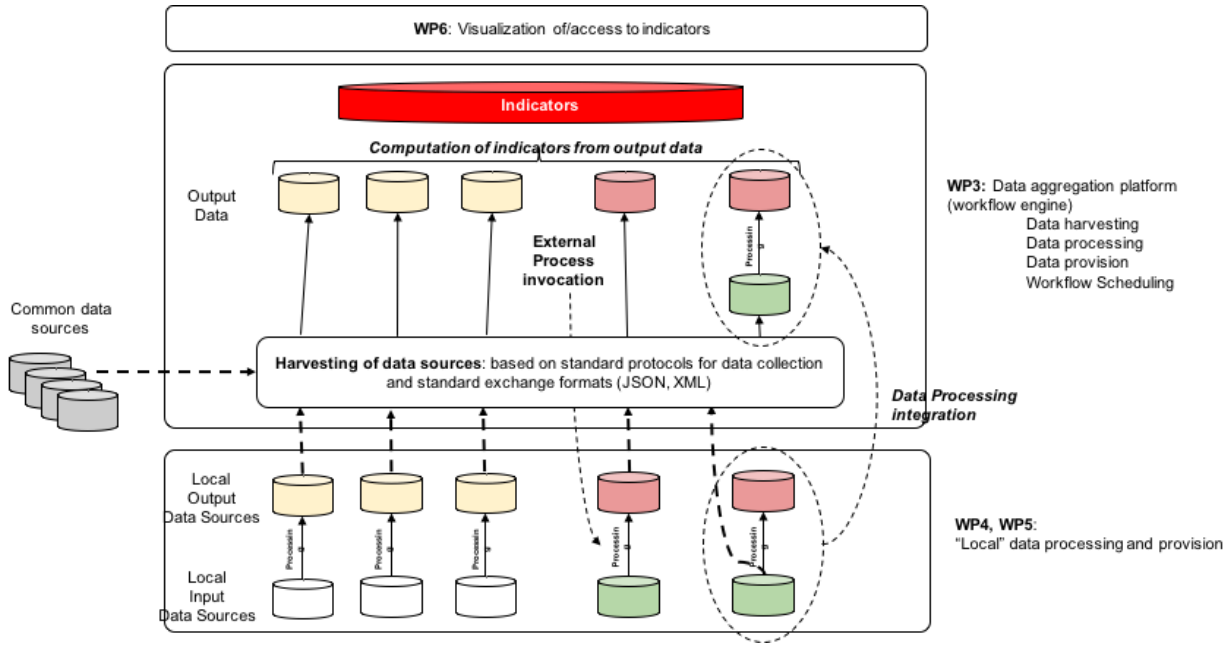

Figure 1 - D4IP workflow integration

As shown in Figure 2, the D4I partners have chosen two out of three integration patterns:

- **Data source integration**: for the **common data sources**. harvesting common data sources and output data sources of a local processing workflow.
- **External workflow integration**: for all workflows with an initial integration of the workflows by collecting their **output data sources.**

The workflows, to be implemented using the D-NET software toolkit, will collect and integrate common information sources and results of data analysis provided in different formats to aggregate them within the D4IP relational database. As we shall database represents a graph where common entities such as funders, projects, topics, and publications are related with all the outcomes of the different analysis workflows.

In this first phase, the interaction with WP6 will be crucial to identify the kind of information required by end-users of the platform when consulting the D4IP database to measure impact of projects and funders. Such analysis will help to understand what kind of content the full-text index will have to contain and how to package it to satisfy the end-users functional expectations. Although some of the indicators may endow intuitive usages, we expect this process to be dynamic and to imply refinements and changes of the full-text index and likely the user interfaces before delivering the final tool.

In the following sections we report on the plan of integration of each partner workflow as described in D3.1, its impact on the data model of the D4IP database, and the functionalities that the platform will implement to deliver an integrated information graph, including entities relative to the common sources and the related results of analysis.

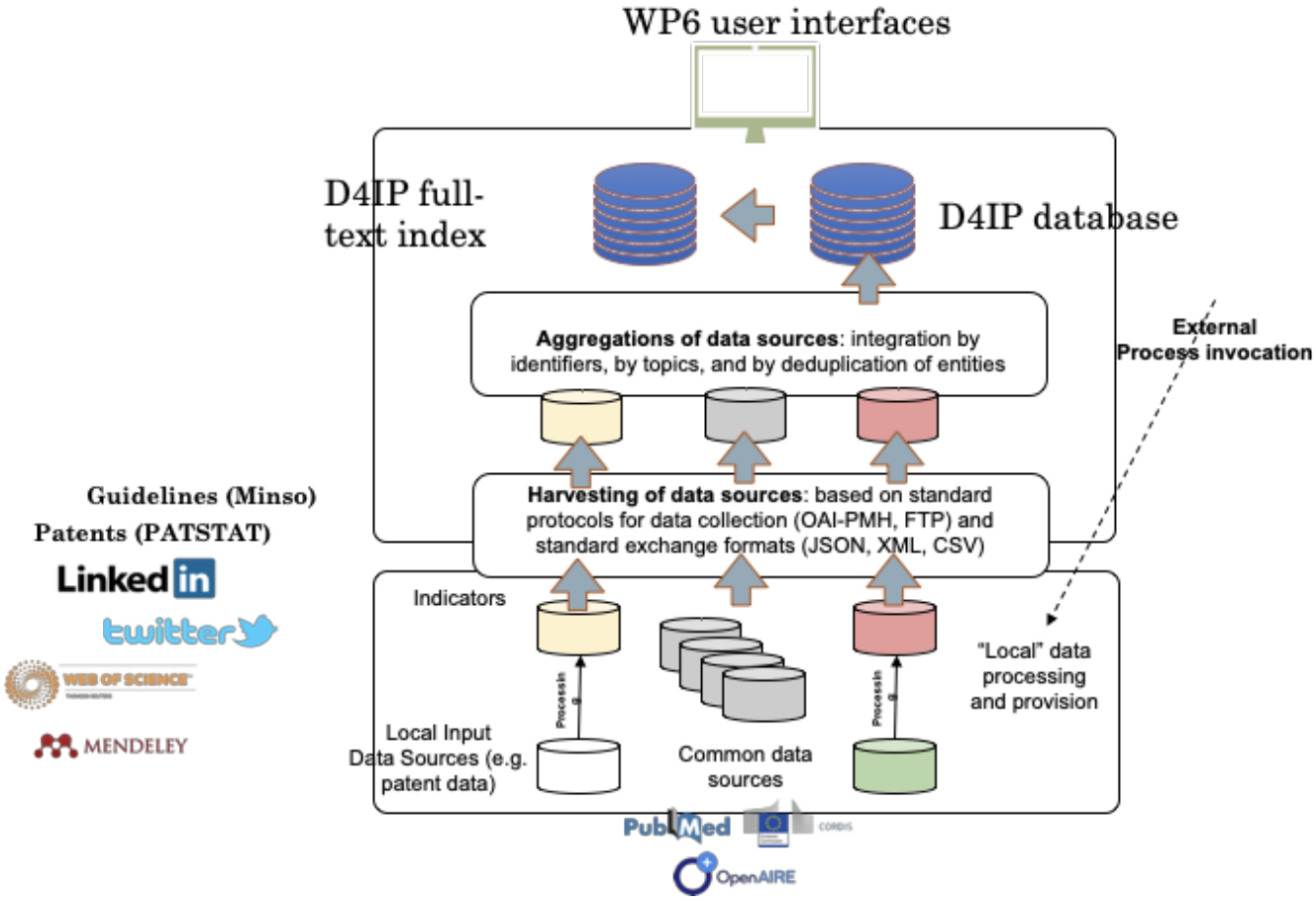

Figure 2 - Final architectural options

2.1 Common sources workflows

From deliverable D3.1 it is clear that the sources to considered across the different use-cases are:

- EC projects: FP7 and H2020, to be collected from OpenAIRE
  - Links to publications in OpenAIRE
- Funders: to be collected from different sources
  - Links to funders as collected from PubMED (collected by PPMI)
  - Links to Swedish projects and funders (collected by PPMI)
- Health-related publications: to be collected from OpenAIRE
  - Links to projects and funders in OpenAIRE

2.2 WP4 workflows

WP4 delivers topic modelling workflows, which are somehow providing a common lingua-franca across the different common sources. In other word, projects, funders and publications will be interrelated by topics, and therefore offer richer input to other data analysis workflow, e.g. interrelation of different indicators by research topic. The workflows operates on the full set of projects, funders, health-related publications, and patents, to return a list of topics related with such entities. The patents documents are collected by ARC from PATSTAT and cannot be shared as a common data source.

The platform will serve WP4 with all the common data sources and then collect the results as an external data source in order to add the links between topics and entities to the DI4P database.

2.3 WP5 workflows

2.3.1 Clinical guidelines analysis

The guideline documents are collected by UoB from Minso Solutions and cannot be shared as a common data source via the D4IP. The analysis workflows perform full-text analysis of clinical guidelines to identify:

1) correlations with funders, project grants, and topics as identified by WP4;
2) calculate indicators of performance of clinical guidelines citation impact, as listed in Table 1.

EC project metadata and topics will be collected from the D4IP. The workflows will return two inputs to be collected by the platform:

Table 1 - Performance of guidelines citation impact indicators

Performance of guidelines citation impact	**Data type**
Project level	[interval values, float]
Programme level	[interval values, float]
Intra-national (affiliation)	[interval values, float]
National level	[interval values, float]
Collaboration at the intra-national and national levels	[interval values, float]

Impact data and links are kept in a relational database that will be queried by SQL calls. The option of collecting a CSV from an FTP site will be considered.

3.3.2 Company data indicators

Workflows of this task deliver indicators of the ability of companies involved in EC projects to deliver innovation. As such the workflow will export both indicators of company innovation (Table 2) and company data (Table 3). Content will be delivered via FTP in the form of JSON files.

Table 2 - Company impact indicators

Column	**Data type**
Company number (PIC number from Cordis)	Integer
Company Innovation Score	Integer
Tangible Innovation count: Pre-market Stage	Integer
Tangible Innovation count: Market Stage	Integer
Intangible Innovation count: Pre-market Stage	Integer
Intangible Innovation count: Market Stage	Integer
Count of commercialization/valorisation activities	Integer

Table 3 - Company data

Initial Data	Interim Data	Final data (uploaded from PPMI to D4IP)
Company_id (int)	Company_id (int)	Company number (PIC number from Cordis) (int)
Site_url (varchar)	Site_url (varchar)	Company Innovation Score (int)
Site_text (text)	Innovation_mention (bool)	Tangible Innovation count: Pre-market Stage (int)
Text_language (varchar)	Tangible_pre_market (bool)	Tangible Innovation count: Market Stage (int)
Extraction_date (date-time)	Tangible_market (bool)	Intangible Innovation count: Pre-market Stage (int)
	Intangible_pre_market (bool)	Intangible Innovation count: Market Stage
	Intangible_market (bool)	Count of commercialization/valorisation activities (int)
	Commercialization_evidence (bool)	

2.3.3 EC project portfolios

Project portfolios will be exposed via REST APIs[1] supported by dedicated data repository system, entitled PALOMAR. PALOMAR Data Analysis and Modelling Platform is an innovative, automated data platform, which incorporates the datasets collected/generated and enrich them with metadata automatically produced by reliable analytics workflows, several scientific instruments as well as key insights and KPIs quantified by modelling tools. PALOMAR is used in D4I to produce project portfolios, which consists of packages of project-related information as described in Table 4.

Table 4 - Project portfolio

Information
The actual Call e-booklet
Project Description along with metadata concerning participants, funding etc.
Project Documentation (Final reports, brief results etc.)
Project Publications

[1] REST APIs access is restricted to the D4IP services

Project Patents
Policy Documents.

Project Portfolios are generated from various sources of raw data and metadata: CORDIS, Pubmed publications, Patents, project reports. The Data4Impact Text Mining workflows consume these data and process them in order to extract useful metadata and insights about projects. In particular:

- **Insights**: record corresponding to a summary of the project portfolio, including information about projects (keywords, disciplines, etc.) and their relations with impact factors (e.g. keywords under innovation context)
- **Indicators**: record that contains measures relative to the portfolio, from input data (e.g. number of publications). Indicators were splitted in four major blocks: admin, scientific, economic and societal where each block encompasses the relevant KPIs and their values (this is still work in progress).

The D4IP, from the end of November 2018, will start collecting the portfolios, the insights, and the metrics, with relationships to the projects, from API.

2.3.4 Project mentions

For each EU-funded project a dataset will be produced that contains the *mentions* per medium regarding the project: news articles, blog posts, fora posts, tweets. Out out of the mentions, the workflow will generate for each project the information in Table 5.

Table 5 - Project mentions

Information
Total Buzz: number of total mentions
Sentiment Analysis: 3 numbers (% percentages) for positive/neutral/negative
Project Source impact measurement: average global rank (in case of news sites, blogs), number of followers (tweets) in search results per query

2.3.5 Social media/metrics

The workflow will analyse how articles linked to EU-funded projects have been tweeted and retweeted. The propagation effect allows to produce performance of such articles based on the conversations taking place around such articles, hence around the related projects. The output of such process is a set of tweeter conversations related to a project, which can offer indicators like the ones listed in Table 4. Such indicators are being currently examined and will be made available in 2019.

3. D4IP database data model

By observing the data analysis workflows above, we have designed an initial relational data model for the D4IP database, where all metadata and indicators about the involved entities will be aggregated. Firstly, we have identified a set of core entities (so-called "throughput data") around which the analysis are spinning, namely projects, publications, and topics. Such entities represent the backbone of the project, becoming the interconnection between the subject of the analysis, i.e. funders and projects, and all the findings and measures related with them. Secondly, we have introduced entities that are pivotal for the local data analysis workflows, required to deliver structured information related with core entities, namely company, guidelines, and patents. Finally, we have identified entities and properties describing the indicators as measured against core and secondary entities.

3.1 Core entities

Many of the questions that the platform will be called to answer will pivot around the funding programmes, the projects, and the topics. This suggests to put such entities at the centre of the data model and arrange for all indicators to be linked to them. The high-level relational schema is depicted in Figure 3.

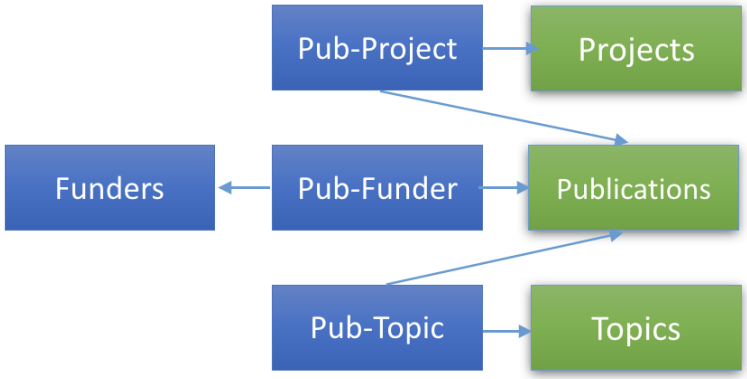

Figure 3 - Core data model entities

3.2 Secondary core entities

The other entities introduced by the data analysis workflows are the following and extend the data model as depicted in figure 4. A set of entities are relative to secondary core entities, introduced by partners to extract indicators:

- Clinical guidelines: linked to publications and topics
- Patents: linked to topics
- Companies: linked to projects and topics

Properties for such elements will be limited to identifier, title, authors, and dates. The rest of the information, such as the text, cannot be shared among partners.

3.3 Indicator core entities

The final and key set of entities is introduced to include the project (and company) indicators resulting from workflow analysis:

- Guidelines citation indicators (detailed properties as indicated in Table 1): linked to guidelines
- Company indicators (detailed properties as indicated in Table 2): linked to companies
- Project portfolios (detailed properties as indicated in Table 4): linked to projects
- Health-related project mention indicators (detailed properties as indicated in Table 5): linked to projects
- Twitter-related project indicators (Table 5): linked to projects

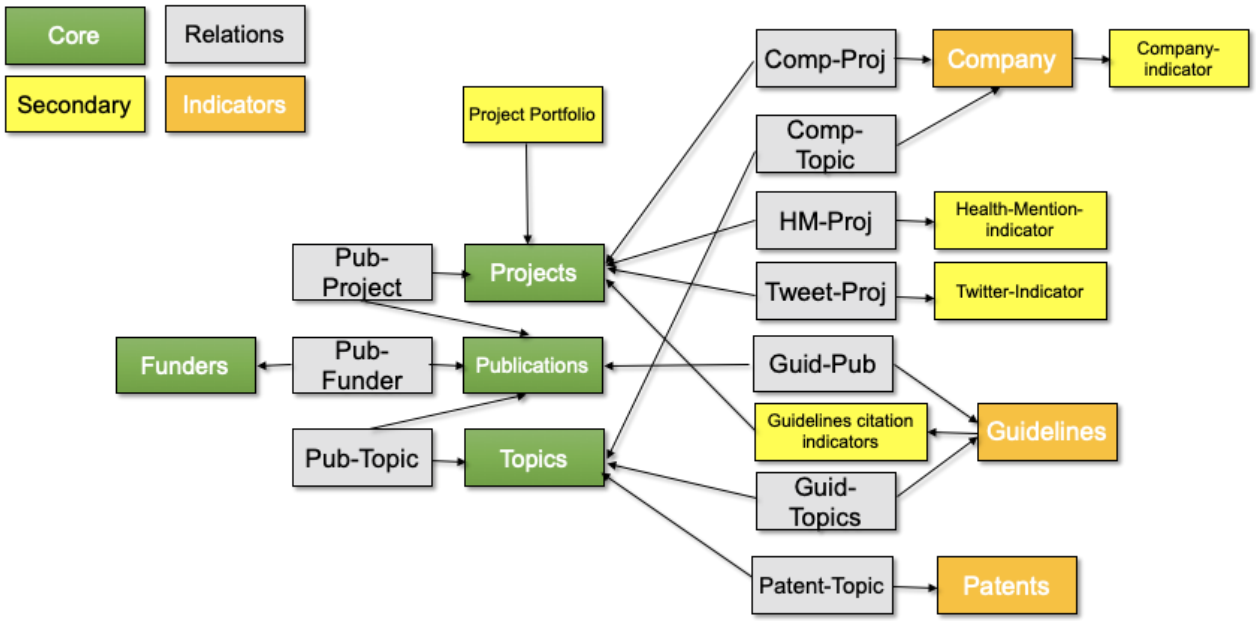

Figure 4 - Derived model entities

From an analysis and discussion with partners, it appeared that different workflows may make use of different identifiers for publications (e.g. DOI, PMCID), projects, and funders (from CORDIS, from EuropePMC, from Swedish ministries databases). Once aggregating the different sources, the D4IP will also perform de-duplication of different manifestations for different entities in order to deliver an integrated graph of concepts.