



Evidence Based Big Data Benchmarking to Improve Business Performance

D5.2 Final Evaluation of DataBench Metrics

Abstract

This deliverable presents the final set of DataBench indicators and introduces the DataBench popularity index – a quantitative measure for ranking the concepts and topics in the area of Artificial Intelligence and Big Data. The results of this deliverable include the report on technical and business benchmark indicators and metrics, as well as quality metrics of DataBench Toolbox. Furthermore, a DataBench Observatory tool for observing the popularity, importance and the visibility of topic terms is announced, developed and discussed in this deliverable.



Deliverable 5.2	Final Evaluation of DataBench Metrics
Work package	WP5
Task	5.1
Due date	31/12/2020
Submission date	26/05/2021
Deliverable lead	JSI
Version	2.1
Authors	JSI (Marko Grobelnik, Inna Novalija, M. Beshir Massri) SINTEF (Arne Berre) LEAD (Todor Ivanov) ATOS (Tomás Pariente)
Reviewers	Nuria de Lama (ATOS), Cristina Pepato (IDC)

Keywords

Big Data Validation, Performance Metrics, Knowledge Graphs, Ontology

Disclaimer

This document reflects the authors view only. The European Commission is not responsible for any use that may be made of the information this document contains.

Copyright Notice

Copyright belongs to the authors of this document. Use of any materials from this document should be referenced and is at the user's own risk.

Contents

Contents	3
Executive Summary	6
1. Introduction	7
2. Objectives.....	8
3. Indicators and Metrics	8
3.1. Technical Benchmark Indicators and Metrics.....	10
3.2. Industry and Business Indicators and Metrics	19
3.3. Quality Metrics of the DataBench Toolbox	22
4. DataBench Observatory Methodology.....	23
4.1. DataBench Observatory Metrics.....	24
4.2. DataBench Popularity Index	25
4.3. Methodology Pipeline	26
4.4. Data Sources	28
4.5. Data Dimensions and Data Formats.....	35
4.6. Pillars.....	35
4.7. DataBench Ontology and Knowledge Graph Monitoring.....	36
5. DataBench Observatory Implementation.....	38
5.1. Frontend and Backend	38
5.2. DataBench Observatory Documentation.....	40
5.3. DataBench Observatory Usage Scenarios and Metrics Results	43
5.4. Integration into DataBench Toolbox.....	48
Summary.....	49
Bibliography	49
Annex: DataBench Ontology Formalization	51

Table of Figures

Figure 1: DataBench Indicators Ecosystem.....	9
Figure 2: DataBench Toolbox Search by Indicators	10
Figure 3: Search by Benchmark-specific Features in the DataBench Toolbox	11
Figure 4: Search by Platform and Architecture Features in the DataBench Toolbox.....	13
Figure 5: Search by Big Data Application Features in the DataBench Toolbox.....	15
Figure 6: Search by Toolbox-specific Features in the DataBench Toolbox.....	17
Figure 7: Technical User Examples for Searches.....	19
Figure 8: Business KPI Benchmarks Overview	20
Figure 9: Search by Business Features in the DataBench Toolbox	21
Figure 10: Search by Use Case in the DataBench Toolbox.....	21
Figure 11: Business User Example for Guided Knowledge Nuggets Search.....	22
Figure 12: Quality Metrics Dashboards	23
Figure 13: DataBench Methodological Framework	24
Figure 14: DataBench Index Components	27
Figure 15: DataBench Methodology Aspects.....	28
Figure 16: Pipeline Challenges	28
Figure 17: DataBench Ontology Snapshot	38
Figure 18: DataBench Observatory (Explanatory menu, accessed in May 2021)	40
Figure 19: DataBench Observatory (Information page, accessed in May 2021)	40
Figure 20: DataBench Observatory (Data sources, accessed in May 2021)	41
Figure 21: DataBench Observatory (Ontology description, accessed in May 2021)	41
Figure 22: DataBench Observatory (Methodology description, accessed in May 2021).....	42
Figure 23: DataBench Observatory (Data distributions, accessed in May 2021).....	42
Figure 24: DataBench Popularity Index (Topics, accessed in November 2020).....	43
Figure 25: Time Series (Topics, accessed in November 2020).....	43
Figure 26: DataBench Popularity Index (November 2020)	44
Figure 27: Time Series (Tools and Technologies, accessed in November 2020)	44
Figure 28: Topic Evolution Visualization (Indicators, accessed in November 2020).....	45
Figure 29: Live News (accessed in November 2020)	45
Figure 30: DataBench Observatory in DataBench Toolbox.....	48

Table of Tables

Table 1: Results from the Benchmark-specific Features in the DataBench Toolbox	12
Table 2: Platform and Architecture Features in Toolbox.....	14
Table 3: Results from the Big Data Application Features in the DataBench Toolbox.....	16
Table 4: Results from the Toolbox-specific Features in the DataBench Toolbox.....	18
Table 5 DataBench Observatory Metrics.....	25
Table 6 Data Sources Format	35
Table 7 DataBench Observatory Metrics Results.....	48

Executive Summary

DataBench Deliverable 5.2 provides the final evaluation of DataBench metrics, including technical and business benchmark indicators and metrics, as well as quality metrics of DataBench Toolbox.

The deliverable revisits the DataBench indicators and their importance in the different search functionalities implemented in the DataBench Toolbox. It demonstrates how the various indicators can serve the different Toolbox user profiles to find the most appropriate benchmarks and Knowledge Nuggets that they are looking for.

The deliverable introduces the DataBench Popularity Index – a quantitative measure for ranking the concepts and topics in the area of Artificial Intelligence and Big Data, and metrics related to the index.

The document presents an overview of the methodology and implementation for the DataBench Observatory - a tool for observing the popularity, importance and the visibility of terms by topic. Particular attention in this tool is dedicated to the concepts, methods, tools and technologies in the area of benchmarking.

In addition, the deliverable discusses the knowledge formalization aspects and presents the DataBench ontology for sharing knowledge about topics, tools and technologies popularity.

1. Introduction

The Final evaluation of DataBench metrics presents a final set of indicators for comparing different datasets. This deliverable illustrates the **DataBench Ecosystem of Indicators** that include relevant indicators and their implementation in the DataBench Toolbox.

The methodology of the **DataBench Index** and implementation of the novel **DataBench Observatory tool** are explained in the current document.

In particular, we provide a description of metrics linked to the DataBench Observatory tool, such as DataBench Index per topic/tool/technology, Number of research papers per topic/tool/technology, Number of jobs per topic/tool/technology, Number of Cordis EU projects per topic/tool/technology, Number of GitHub projects per topic/tool/technology, General popularity of topic/tool/technology.

The deliverable discusses the data sources behind the DataBench Index and aspects related to the DataBench ontology. Scenarios for several targeted user groups are provided and illustrated with examples.

The document is structured as follows:

- Section 1 provides the introduction to the deliverable.
- Section 2 describes the objectives and the work performed.
- Section 3 presents the relevant indicators, followed by quality metrics in DataBench Toolbox.
- Section 4 introduces the DataBench Observatory tool, with methodology.
- Section 5 provides the implementation description of the DataBench Observatory tool.
- Summary section provides the conclusions of the document and
- The Annex contain the formalization of the DataBench ontology.

Deliverable 5.2 is following DataBench Deliverable 5.1 – Initial Evaluation of DataBench Metrics [4]. The current deliverable is related to the DataBench Deliverables 5.4 – Analytic modelling relationships between metrics, data and project methodologies [6] and 5.5 – Final report on methodology for evaluation of industrial analytic projects scenarios [7]. In particular, Deliverable 5.4 discusses DataBench metrics and references the DataBench Observatory tool and its application for the benchmarking analysis. Deliverable 5.5 provides a detailed analysis of Big Data projects with respect to DataBench results and a scalable overview of Cordis EU projects within the DataBench Observatory.

2. Objectives

One of the objectives set for WP5 is developing a methodology that would allow for producing metrics to make comparable different datasets and provide an understandable metrics landscape.

In order to fulfill this objective, in this document we are looking at the relevant indicators and metrics and their implementation in the DataBench Toolbox.

Furthermore, we have developed a methodology for obtaining the DataBench Index – a quantitative measure for ranking the concepts and topics in the area of Artificial Intelligence and Big Data.

The DataBench Index in the broad sense reflects the popularity, importance and visibility of terms related to AI and Big Data topics across different resources. At the same time, the DataBench Index allows for comparing and ranking tools and technologies specifically related to benchmarking processes.

In such way, the objective of work performed as part of Deliverable 5.2 is to develop a methodology for producing the DataBench Index and the DataBench Observatory tool for implementation of the index via different pillars.

3. Indicators and Metrics

In WP1, we introduced the DataBench Ecosystem of Indicators that include both technical and business and industry relevant indicators. These categories of indicators are also implemented in the DataBench Toolbox and applied in the different Work Packages. [Figure 1: DataBench Indicators Ecosystem](#) depicts the DataBench Indicators Ecosystem. Indicators are divided in two main categories: 1) Business Features and 2) Technical Features, which consist of Big Data Application Features, Platform and Architecture Features and Benchmark-Specific Features. All these feature categories are describing and measuring the various DataBench Framework dimensions, which are different from the above mentioned DataBench Index. They are focused on the DataBench Framework, whereas the DataBench Index is measuring the relevance of DataBench terms based on the evaluation of different data sources.

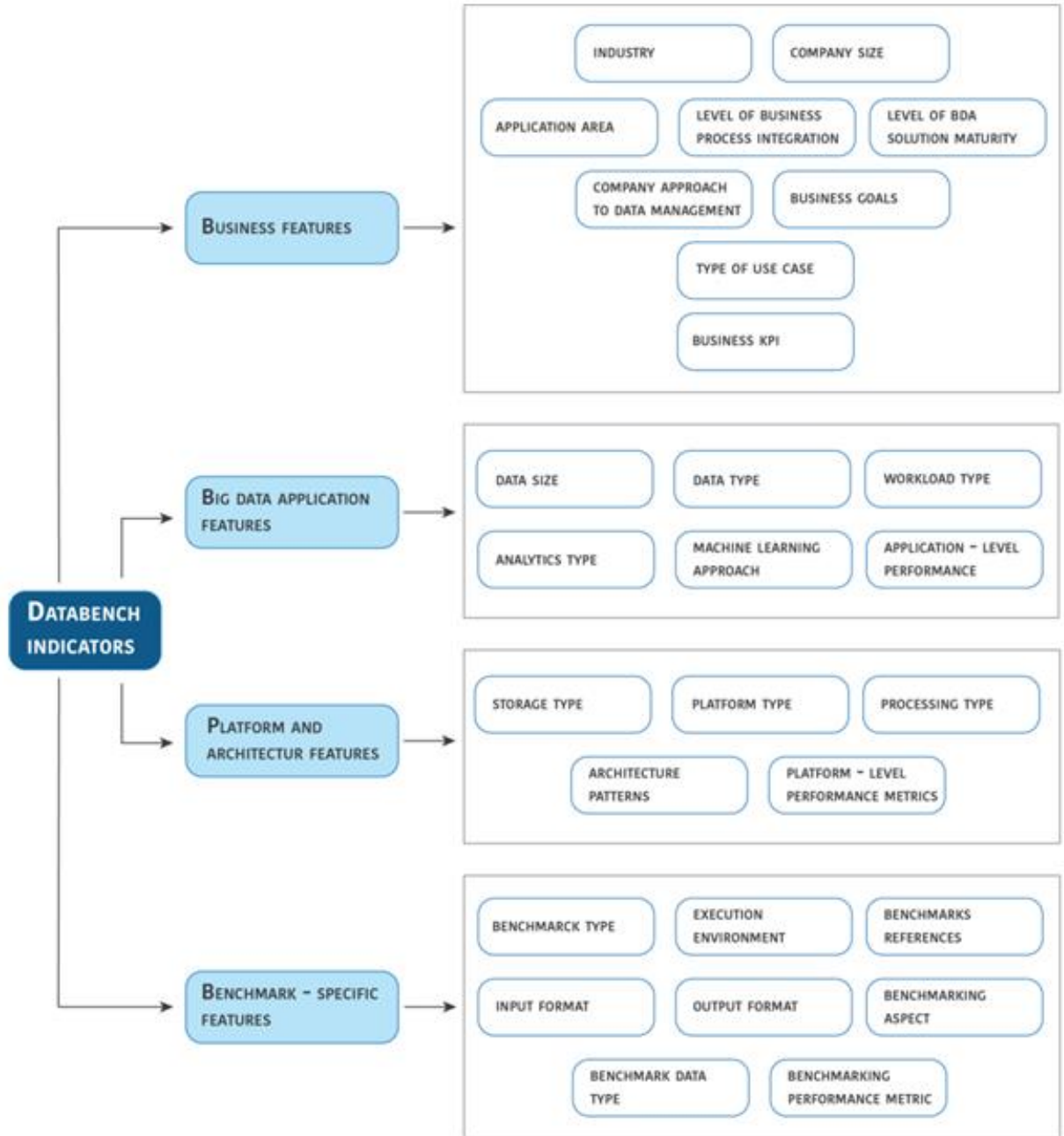


Figure 1: DataBench Indicators Ecosystem

The DataBench Ecosystem of Indicators was introduced in D1.1 [1] with the goal to describe the most common technical and business indicators that are used when describing, comparing and searching for the optimal technologies and benchmarks. Furthermore, the DataBench Framework was presented in D1.2 [2]. It is based on the structure of the BDVA Reference Model [8] that classifies the benchmarks into vertical and horizontal and further relates then to business-oriented benchmarks. Using the DataBench Framework we classified close to 100 AI and Big Data benchmarks and created the DataBench Framework

Matrix also introduced in D1.2. The objective of this matrix is to assist in the benchmark search within the Toolbox by providing as many features to the benchmarks as possible. In this way users with different background knowledge and levels of expertise varying from niche technical experts to business experts will be able to navigate and extract knowledge. Combining the DataBench Framework Matrix and the Ecosystem of Indicators, we extended the abilities to search and compare the different benchmarks according to their features. The resulting approach is now implemented as different searching options in the DataBench Toolbox as depicted on [Figure 2: DataBench Toolbox Search by Indicators](#).






SEARCH BY TAG	
Navigate through the tag categories:	
 Business Features	▼
 Big Data Application Features	▼
 Platform and Architecture Features	▼
 Benchmark-specific Features	▼
 Toolbox-specific Features	▼

Figure 2: DataBench Toolbox Search by Indicators

The implementation details for the Toolbox search functionalities are provided in WP3 Deliverable 3.4 [3]. The focus of this section is to demonstrate the most unique features of the Toolbox, identified in the evaluation phase of the project in Task 5.2, and briefly explain why the different search functionalities are so important.

3.1. Technical Benchmark Indicators and Metrics

The first logical benchmark metrics that we classified were the **benchmarking performance metrics** that each technical benchmark reports after execution. Some examples for such metrics are execution times or latency, throughput, cost of the system under test, energy consumption and many more. The purpose of these metrics is to enable the comparison between the tested software or hardware systems, specific components that they stress test or the complete system including all components. This category of metrics depends on the type of benchmark in which they are reported, as some benchmarks focus on micro operations, while others are more complex and simulate a complete application workload. Furthermore, as the AI and Big Data fields are evolving and new emerging technologies are being developed, so is the need for new benchmarks and

technical metrics that more precisely measure differences in the new software and hardware components. For example, new MLPerf metrics such as “the time required to train a model on the specified dataset to achieve the specified quality target” [9] and “scenario specific metrics” [10]. These types of new metrics are more complex and involve many more aspects describing the hardware and software components of the system under test as well as the specific datasets and different processing steps in which they are involved. This is in alignment to our DataBench Framework indicators depicted in Figure 1: DataBench Indicators Ecosystem. In particular, the **Benchmark-specific Features** extend the **benchmarking performance metrics** with **benchmark data type**, **execution environment**, **benchmarking aspects**, **benchmarking type** and **input/output data formats** that describe the technical benchmarks in greater detail and provide valuable information for both searching and comparing benchmarks. The definition of each indicator is available in D1.1. Figure 3: Search by Benchmark-specific Features in the DataBench Toolbox shows how these indicators are implemented in the Toolbox and the most common values represented as tags. The number next to each tag shows the number of benchmarks that relate to this indicator value. The bigger the number is, the more general is the indicator value, which will require to use further indicators to identify the most suitable benchmark for the required purpose.

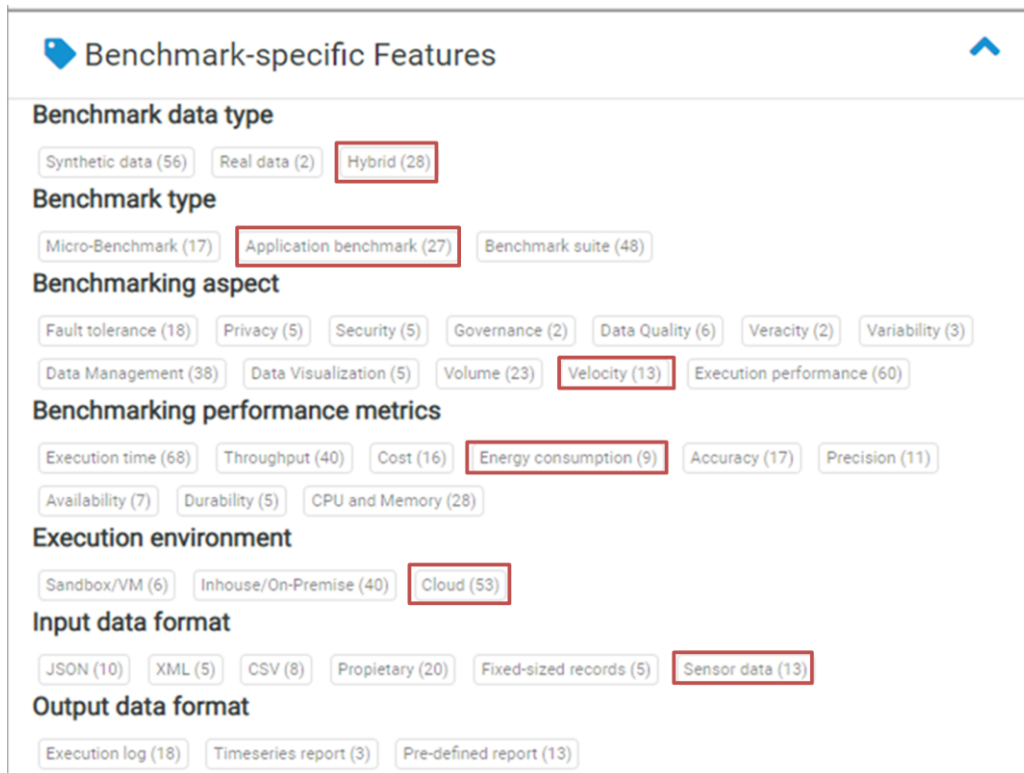


Figure 3: Search by Benchmark-specific Features in the DataBench Toolbox

Table 1 shows the resulting technical benchmarks and knowledge nuggets from an example Guided Benchmark search based on Benchmark-specific Feature tags. In yellow are marked the selected search tags, in blue are the resulting top 10 technical benchmarks and in orange are in this case the top 4 knowledge nuggets. In this search for Benchmark-specific Features were picked tags for each of the indicator categories to illustrate the variety of technical features describing the benchmarks.

Go to Search --> Guided Benchmark Search and select the Benchmark-specific Features drop-down menu. Then click on the following tags:					
Hybrid	Application benchmark	Velocity	Energy Consumption	Cloud	Sensor Data
The search returns the following list of benchmarks (only the top 10):					
BigBench V2	BigBench V2	Yahoo Streaming Benchmark (YSB)	ABench	HiBench	BenchIoT
HiBench	owperf (CLASS)	AIM Benchmark	AlBench	owperf (CLASS)	CloudSuite
owperf (CLASS)	Yahoo Streaming Benchmark (YSB)	BigDataBench	BenchIoT	Yahoo Streaming Benchmark (YSB)	IoTAbench
ABench	AdBench	BigFUN	Benchip	Yahoo! Cloud Serving Benchmark (YCSB)	IoT Bench
AIMatrix	AIM Benchmark	CloudSuite	BigDataBench	ABench	RIOTBench
BenchIoT	BigFrame	Framework of Load & Integration for Cloud Pub/Sub (FLIC)	IoT Bench	AIM Benchmark	Senska
Benchip	CityBench	IDEBench		ALOJA	StreamBench
BigBench	Graphalytics	Linear Road		AMP Lab Big Data Benchmark	TPCx-IoT
BigDataBench	IDEBench	Stream WatDiv		Benchip	VisualRoad
BlockBench	Linear Road	StreamBench		BigDataBench	
The search returns the following list of knowledge nuggets (only the top 4):					
Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Agriculture Architectural Blueprints
	Benchmarks under Evaluation for the DataBench Toolbox	Telecommunications - Network resource and capacity optimization Architectural Blueprints	Standard Performance Evaluation Corporation (SPEC)		Benchmarks features matrix
	Project ALOJA		Transaction Processing Performance Council		IoT Ingestion and Authentication
	Technical Benchmarks: Definition and Taxonomy				IoT Pipeline pattern

Table 1: Results from the Benchmark-specific Features in the DataBench Toolbox

With the growing complexity of new systems and Big Data and AI technologies, even more specific benchmarks are emerging. This makes it even harder for practitioners to keep the

overview of all of them and pick the one measuring and reporting the best metrics for their requirements. To enable this, the only viable solution is to provide as much information for each benchmark and its environment as possible. Therefore, we introduced two additional groups of features: Platform and Architecture Features and Big Data and AI Application Features. The Platform and Architecture Features search is depicted in Figure 4 and consists of 5 indicators which describe the typical features of today's AI and Big Data systems. With red are marked tags selected for a sample search which is described in the following table below.

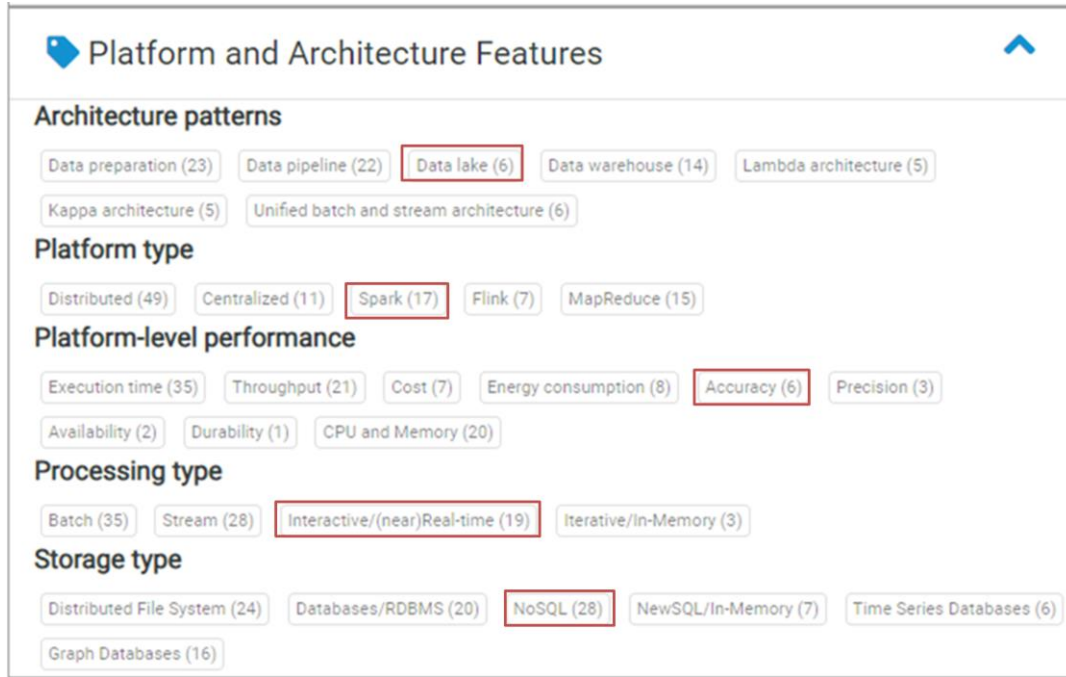


Figure 4: Search by Platform and Architecture Features in the DataBench Toolbox

Table 2 demonstrates how the results based on a Guided Benchmark search filtering the Platform and Architecture Features tags looks like. Again, in this case are listed the top 10 technical benchmarks in blue and the top 5 knowledge nuggets in orange. It is obvious that the number of results (benchmarks and nuggets) vary for the different tags. In both cases one has to look at as many tags as possible and then cross select them to obtain the results that apply to all relevant tags. For example, there are many benchmarks and nuggets for the NoSQL category, whereas there are very few results for the Accuracy metric, which means that that it will be important tag in the selection criteria.

Go to Search --> Guided Benchmark Search and select the Platform and Architecture Features drop-down menu. Then click on the following tags:				
Data Lake	Spark	Accuracy	Interactive/(near)Real-time	NoSQL
The search returns the following list of benchmarks (only the top 10):				
Berlin SPARQL Benchmark (BSBM)	BigBench V2	AlBench	Yahoo Streaming Benchmark (YSB)	Yahoo Streaming Benchmark (YSB)

CALDA	HiBench	HPC AI500	Yahoo! Cloud Serving Benchmark (YCSB)	Yahoo! Cloud Serving Benchmark (YCSB)
Hadoop Workload Examples	ABench	Linear Road	AMP Lab Big Data Benchmark	AMP Lab Big Data Benchmark
SWIM	BigDataBench	MLBench	Berlin SPARQL Benchmark (BSBM)	BigBench
	PigMix	Training Benchmark for DNNs (TBD)	BigFUN	BigDataBench
	PRIMEBALL		CBench-Dynamo	BigFUN
	Sanzu		CityBench	CBench-Dynamo
	SparkAIBench		Graphalytics	CloudRank-D
	SparkBench		IDEBench	CloudSuite
	StreamBench		LinkBench	GDPRBench
The search returns the following list of knowledge nuggets (only the top 5):				
Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	NoSQL - Key-Value DB
Data Lake	Data features Star Diagram		Classification of data processing architectures	NoSQL - Document DB
Use case independent blueprints - Data Management Systems - Real-time Stream Storage Architecture	Practical example of creating a blueprint and derived cost-effectiveness analysis: Targeting the Telecommunications Industry		Data features Star Diagram	NoSQL - Graph DB
Use case independent blueprints - Data Management Systems - Real-time File Storage Architecture	Project ALOJA			NoSQL - Time Series DB
				NoSQL - Wide column DB

Table 2: Platform and Architecture Features in Toolbox

The Big Data and AI Features search is depicted in Figure 5 and consists of 7 indicators which are mainly taken from the BDVA reference model. Here again are selected only six tags marked with red to demonstrate the search results in the following table.

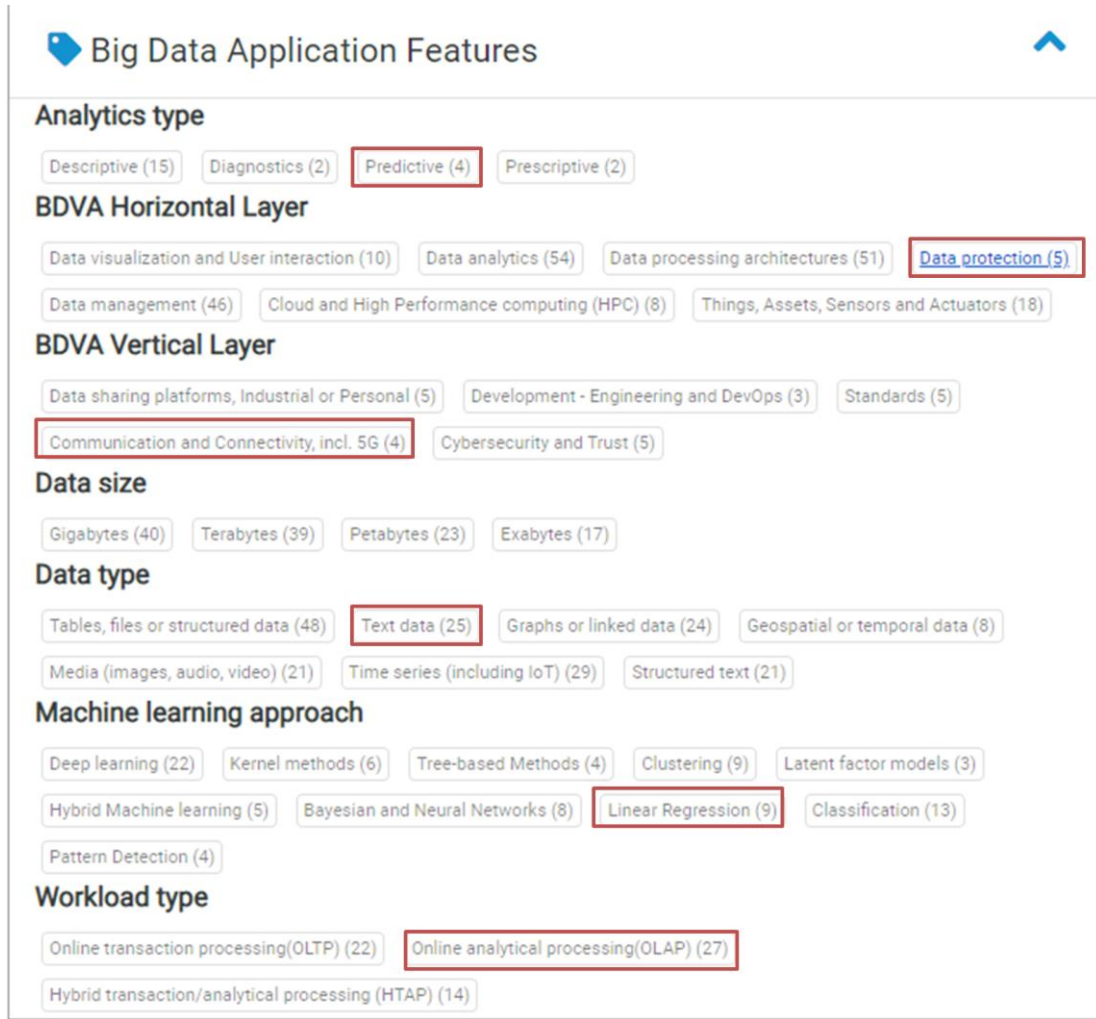


Figure 5: Search by Big Data Application Features in the DataBench Toolbox

Table 3 shows the resulting technical benchmarks and knowledge nuggets from the tags selected in the Guided Benchmark search for Big Data Application Features on Figure 5. In this case were selected on purpose tags that are not that common to demonstrate that still can be applied to get useful results. As the Toolbox matures more new benchmarks and other relevant resources will be added to extend its coverage, especially in the field of AI and machine learning.

Go to Search --> Guided Benchmark Search and select the Big Data Application Features drop-down menu. Then click on the following tags:					
Predictive	Data protection	Communication and Connectivity	Text data	Linear Regression	Online analytical processing (OLAP)
The search returns the following list of benchmarks (only the top 10):					
RIoT Bench	ALOJA	ABench	BigBench V2	HiBench	BigBench V2
Senska	GDPRBench	NNBench-X	HiBench	CloudRank-D	HiBench
	HERMIT		AlBench	MLBench	AIM Benchmark

	TERMinator Suite		AIMatrix	OpenML Benchmark Suites	AlotBench
			BigDataBench	PUMA Benchmark Suite	Benchip
			CloudRank-D	Sanzu	BigBench
			CloudSuite	SparkAIBench	CALDA
			DAWNBench	SparkBench	Deep Learning Benchmarking Suite (DLBS)
			DeepMark (Convnet)		DeepBench
			Edge AIBench		Graphalytics
The search returns the following list of knowledge nuggets (only the top 5):					
Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix	Benchmarks features matrix
Data features Star Diagram		Technical benchmarks Star Diagram	Data features Star Diagram		AI&ML Development Platforms (IDEs)
Evidence of business process performance - Yield prediction use case (Agriculture)			DataBench Generic Data Pipeline (4 steps data value chain)		AI&ML Frameworks
Use case independent blueprints - Data processing and exploitation systems - Data visualization and business intelligence architecture			Natural Language Processing (NLP)		AI&ML Libraries
			Technical benchmarks Star Diagram		AI&ML Platforms

Table 3: Results from the Big Data Application Features in the DataBench Toolbox

The Toolbox-specific Features shown on Figure 6 is an additional group of indicators that classify the knowledge and materials available in the Toolbox in tags like the technical benchmark and platform features but representing more abstract and general terminology. This means that the results they return can be either benchmarks and knowledge nuggets

or only one of both. Again, with red are marked some example tags that will be further explored below in the table with results.

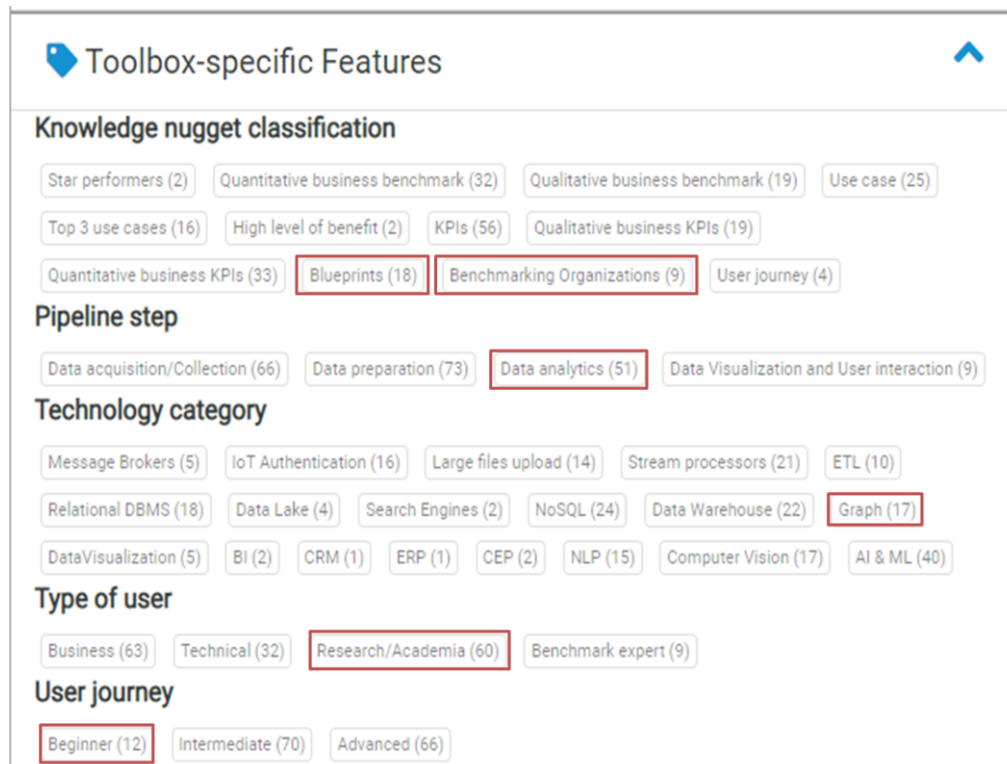


Figure 6: Search by Toolbox-specific Features in the DataBench Toolbox

Table 4 shows the results from the above selected tags for Guided Benchmark search with the Toolbox-specific Features. With yellow are marked the search tags. Their results are marked with orange when they result in the top 10 knowledge nuggets or marked with blue when the results are the top 10 technical benchmarks. The reason for this is that some search tags are technical and related more to the benchmarks, while others are more general, related to architectural blueprints or the usage of the Toolbox like the type of users or the user journey. In general, the goal is to offer the end-users easier way to navigate and enable them to quickly discover the different categories of information and knowledge available in the DataBench Toolbox.

Go to Search --> Guided Benchmark Search and select the Toolbox-specific Features drop-down menu. Then click on the following tags:					
Blueprints	Benchmark Organizations	Data Analytics	Graph	Research/ Academia	Beginner
The search returns the following list of benchmarks or knowledge nuggets (only the top 10):					
Linked Data Integration and Publication Pipeline pattern	BDVA TF6 SG7: Big Data Benchmarking Sub Group	BigBench V2	HiBench	Quantitative Benchmarks for the Top 3 Use Cases in Utilities, Oil and Gas	8 KPIs to Measure Big Data Business Impacts - From KPIs to Benchmarks
Agriculture Architectural Blueprints	BenchCouncil	HiBench	Yahoo! Cloud Serving Benchmark (YCSB)	8 KPIs to Measure Big Data Business Impacts	Benchmark-specific Features

				- From KPIs to Benchmarks	
DataBench Generic Data Pipeline (4 steps data value chain)	Benchmarking communities	AdBench	BigDataBench	Agriculture Architectural Blueprints	Benchmarking communities
Earth Observation and Geospatial Pipeline pattern	Hobbit platform and community	AlBench	CityBench	Benchmarks features matrix	Benchmarks under Evaluation for the DataBench Toolbox
Financial services - Fraud detection Architectural Blueprints	Linked Data Benchmark Council (LDBC)	AIM Benchmark	GARDENIA	Financial services - Fraud detection Architectural Blueprints	Big Data Application Features
Generic Big Data Analytics Blueprint and mappings to standards	MLPerf Community	AIMatrix	gMark	Healthcare - Patient monitoring Blueprint	Classification of data processing architectures
Genomic Pipeline pattern	Securities Technology Analysis Center (STAC)	AlotBench	Graphalytics	Project ALOJA	Data features Star Diagram
Healthcare - Patient monitoring Blueprint	Standard Performance Evaluation Corporation (SPEC)	ALOJA	Hobbit Benchmark	Qualitative Benchmarks by Company Size	DataBench Generic Data Pipeline (4 steps data value chain)
IoT Pipeline pattern	Transaction Processing Performance Council	Benchip	LinkBench	Qualitative Benchmarks by Industry	Platform and Architecture Features
Mapping technologies on architectural blueprints		BigBench	MidBench	Qualitative Benchmarks for Large Enterprises	Technical benchmarks Star Diagram

Table 4: Results from the Toolbox-specific Features in the DataBench Toolbox

Finally, Figure 7 summarizes with visual examples the most unique search functionalities of the DataBench Toolbox that were described in this section. By asking a typical technical question as part of the development and optimization of AI and Big Data systems, the practitioners can search for specific benchmark, technology, framework or best practice architecture pattern in the Toolbox. By filtering according to the most relevant tags and cross selecting the results from the multiple searches they can obtain valuable technical advices in various industry specific use cases.

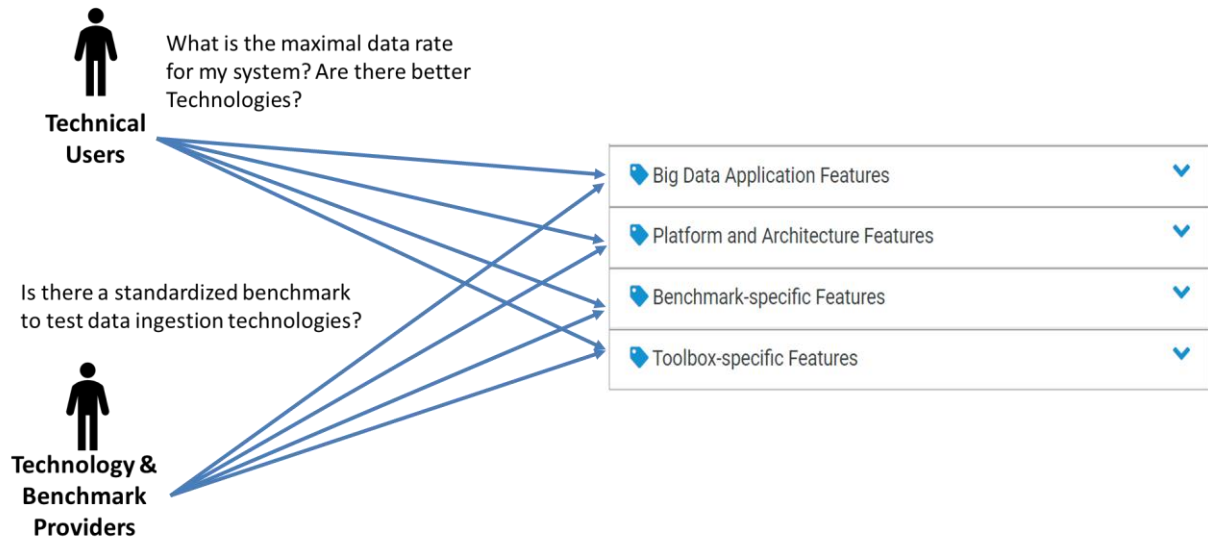


Figure 7: Technical User Examples for Searches

3.2. Industry and Business Indicators and Metrics

Along with the Technical Indicators, described in detail above, there are Business Indicators that were also introduced in D1.1 and depicted on Figure 1. The Business Features can be divided into four main subgroups:

1. The classification of business users – industry and company size
2. The type of BDA implementation – application area, level of business process integration, level of BDA solutions maturity, company approach to data management, and main business goals
3. The type of use case – cross-industry and industry-specific
4. Business impact KPIs, which correspond to industrial benchmarks



They represent different metric indicators described in more detail in D4.4 DataBench Handbook and in the related WP2 and WP4. For example, the 4th subgroup consisting of Business impact KPIs is very important as the KPIs can be used as valid benchmark metrics by researches or business users for each of the industry and company-size segments in which they are measured. The KPIs can be divided into two different categories, according to how they are evaluated (Qualitative) and what they measure (Quantitative). Figure 8 depicts all selected KPIs, their definition and the type of measurement, which can be quantitative or qualitative.

KPI	Definition	Metrics
Revenues increase	Increase in company revenues thanks to the adoption of BDA	Quantitative benchmarks: % increase measured as median of the sample
Profit increase	Increase in company profit thanks to the adoption of BDA	
Cost reduction	Reduction in process costs thanks to the introduction of BDA	
Time efficiency	Efficient use of time in business processes	Qualitative benchmarks: average rating on a scale of 1–5 based on the following ratings: <ul style="list-style-type: none"> • Less than 5% improvement = 1 • 5–9% = 2 • 10–24% = 3 • 25–49% = 4 • 50% or more = 5
Product/Service quality	Product/Service features corresponding to users' implied or stated needs and impacting their satisfaction	
Customer satisfaction	A measure of customers' positive or negative feeling about a product or service compared with their expectations	
New Products/ Services launched	A measure of the number of new products and/or services enabled by data-driven innovation and launched by the company after engaging in the Big Data investment	
Business model innovation	Novel ways of mediating between companies' product and economic value creation (for example, moving from traditional sales to service subscription models)	

Figure 8: Business KPI Benchmarks Overview

Source: DataBench D2.4 Deliverable *Benchmarks of European and Industrial Significance*

Along with search options for technical indicators and metrics, the DataBench Toolbox provides search functionalities for Business Features (illustrated on Figure 9: Search by Business Features in the DataBench Toolbox). The four groups of indicators that are described above together with their categories of tags are shown on Figure 9. A guided search by any of Business Feature tags results in obtaining of a list of knowledge nuggets that can contain both business and technical information. Similar to the technical searches, here again one needs to perform multiple searches based on the most relevant tags and then cross reference the resulting nuggets to identify the most relevant ones. In this case, it is also useful to go through the provided knowledge nuggets as they may contain useful additional information about the relevant topic.


Business Features


Application Area

Customer service and support (1)
Engineering (0)
Research and development (R&D) (1)

Product innovation (new business initiatives) (0)
Maintenance and logistics (0)
Marketing (0)
Finance (1)

HR and legal (0)
Sales (0)
Product management (0)
Governance, risk and compliance (0)
IT and data operations (0)

Business Goals

Better understanding customer behaviour and expectations (1)
Optimise pricing strategies and go-to-market programmes (0)

Product, services or programme improvement and innovation (0)
Improve understanding of the market and competitors (0)

Improve and optimise business processes and operations (0)

Improve facilities, equipment design, maintenance and utilisation (0)
Improve operational, fraud and risk management (0)

Implement better regulatory compliance and financial controls (0)

Business KPI

Increase in the number of products/services launched (16)
Customer satisfaction (15)
Business model innovation (16)

Revenue and profit growth (33)
Product/service quality (17)
Time efficiency (17)
Cost reduction (34)

Company Size

10 to 49 (9)
50 to 249 (9)
250 to 499 (9)
500 to 999 (9)
1000 to 2499 (9)
2500 to 4999 (9)
5000 or more (9)

Industry

Agriculture (9)
Banking, insurance, other financial services (9)
Business or professional services, excluding IT services (7)

IT Services (7)
Healthcare (9)
Manufacturing process (9)
Manufacturing discrete (9)
Retail trade (11)

Wholesale trade (7)
Telecommunications (10)
Media (8)
Transport and logistics (9)
Utilities (7)
Oil & Gas (7)

Level of BDA solutions maturity

Currently using (2)
Piloting or implementing (0)
Considering or evaluating for future use (0)
No use and no plans (0)

Level of business process integration

High (where there is real-time integration with business processes) (0)

Medium (where there are mixed levels of integration with business processes) (0)



Low (where Big Data reports and dashboards are processed in a batch) (0)

Figure 9: Search by Business Features in the DataBench Toolbox

There is another option to Search by Use Case, shown on Figure 10. It consists currently of 12 categories/tags that group multiple technical and business knowledge nuggets, capturing knowledge from all DataBench working packages and deliverables.

SEARCH BY USE CASE

Select use case


Use Case


Use Cases

Price optimization (6)
New product development (7)
Risk exposure assessment (5)
Regulatory intelligence (5)

Customer profiling, targeting, and optimization of offers (8)
Customer scoring and/or churn mitigation (2)
Fraud prevention and detection (4)

Product & Service Recommendation systems (3)
Automated Customer Service (3)
Supply chain optimization (5)
Predictive Maintenance (7)

Inventory and service parts optimization (3)

Figure 10: Search by Use Case in the DataBench Toolbox

In conclusion, Figure 11 illustrates visually the different search options provided by the DataBench Toolbox to a business user or practitioner. Using the specific search features and tags one can quickly start with a question and navigate through the search menus and find a useful technical or business-related information about their field or industry.

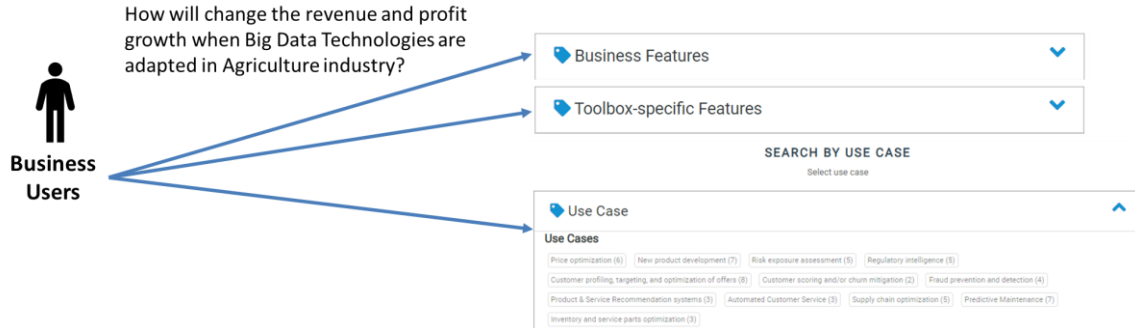


Figure 11: Business User Example for Guided Knowledge Nuggets Search

3.3. Quality Metrics of the DataBench Toolbox

Up to now, all presented DataBench Framework indicators and metrics were focused on either describing the Big Data and AI benchmark features or the features of the environments in which they run and stress test. There is another type of metrics that we call quality characteristics/metrics and are also available in the DataBench Toolbox but report actual statistics for the utilization of the Toolbox platform. The objective of this group of metrics is to dynamically assess the technical usability, relevance, scale and complexity of the DataBench Framework with respect to usage and utilization. These metrics are focused on the Toolbox platform characteristics such as hardware resource utilization, number of benchmark searchers, number of active users, number of implemented/integrated benchmarks and many more. All specification details are defined in D5.3 [5] and later the implementation in the DataBench Toolbox is documented in D3.4 [3]. The metrics are divided in nine categories: 1) Effectiveness; 2) Efficiency; 3) Satisfaction; 4) Performance; 5) Compatibility & Portability; 6) Usability; 7) Reliability; 8) Maintainability and 9) Functional suitability as depicted on Figure 12: Quality Metrics Dashboards. Each category consists of one or more metrics that display visually the metric values. As the DataBench Toolbox gains more users these metrics will be more useful not only for the Toolbox administrators, but also for the technical and business users to backtrack their search history and follow the evolution of the platform.










 Effectiveness	 Efficiency	 Satisfaction
 Performance	 Compatibility/Portability	 Usability
 Reliability	 Maintainability	 Functional suitability

Figure 12: Quality Metrics Dashboards

Source: DataBench D3.4 Deliverable *Release Version of DataBench Toolbox including visualization and search components*

4. DataBench Observatory Methodology

The **DataBench Observatory** is a tool for observing the popularity, importance and the visibility of terms related to Artificial Intelligence and Big Data topics. Particular attention in this tool is dedicated to the concepts, methods, tools and technologies in the area of benchmarking.

The DataBench Observatory is based on the DataBench Popularity Index that is calculated for ranking the terms by topic in time.

Figure 13: DataBench Methodological Framework presents an updated DataBench Methodology Workflow and Implementation, that includes the DataBench Observatory.

The DataBench Observatory methodology and DataBench Index are based on the following aspects:

1. Establishing the relevant data sources and data gathering.
2. Identifying the required dimensions in data and mapping terms from different data sources.
3. Ranking the results from the multiple sources, including normalization of the results. Building the DataBench Index.
4. Identifying point of reference. The starting dates for data, which will affect the relative trendiness. The different data collections vary in time frame – for instance, the jobs data collection contain data that are from two years ago, while research papers data collection contain data that go back in time for over 10 years.
5. Differentiating between topics and subtopics. In particular, identification of the topic terms/tools/technologies related to benchmarking.
6. DataBench Observatory interactive implementation, with visualizations and ranking interface.

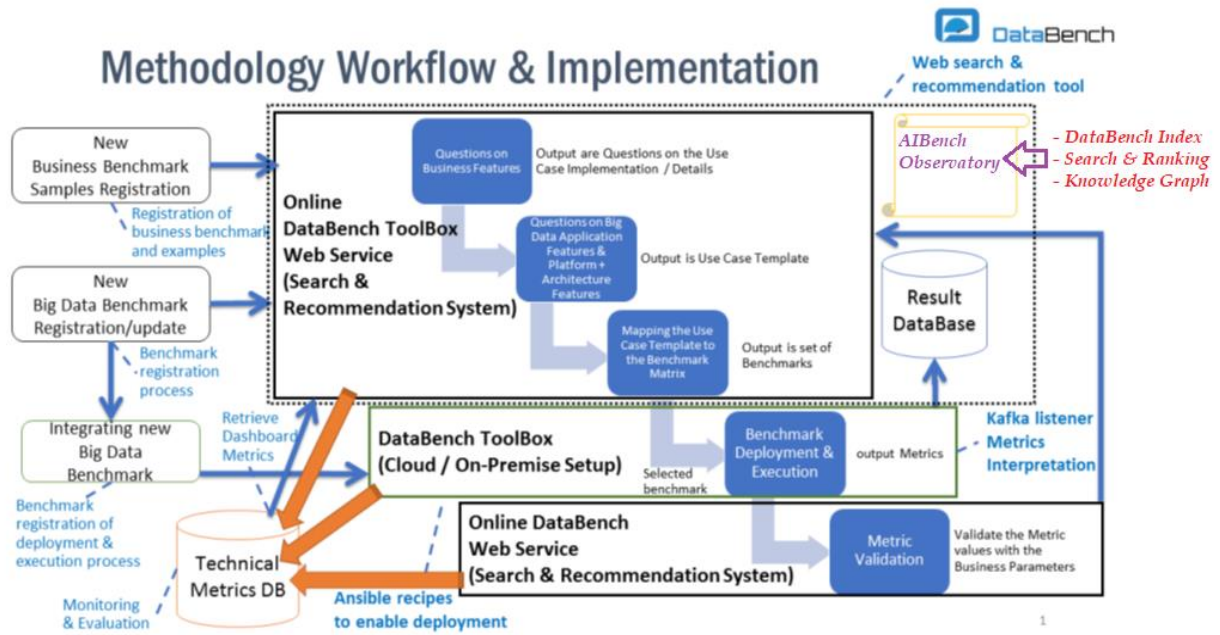


Figure 13: DataBench Methodological Framework

Below in Section 4 we describe the main aspects related to the development of the DataBench Observatory.

4.1. DataBench Observatory Metrics

Below in Table 5 DataBench Observatory Metrics we provide a description of metrics linked to the DataBench Observatory tool.

METRIC NAME	METRIC DESCRIPTION
DataBench index per topic	The metric illustrates the popularity and visibility of the specific topic (topic from the area of Artificial Intelligence, Big Data with particular attention to Benchmarking) on a monthly basis and overall. The primary metric objective is to provide the user with highly important and emerging topics based on combined research and development, industry and general interest data.
Number of research papers per topic	The metric displays the topic popularity in the research domain on a monthly level and overall. The metric presents highly important and emerging research topics.
Number of jobs per topic	The metric displays the topic popularity in the industry/production domain on a monthly basis and overall. The metric presents highly important and emerging topics in industry/production.

Number of Cordis EU projects per topic	The metric displays the topic visibility among EU research and development projects. The metric combines both research and production popularity aspects.
Number of GitHub projects per topic	The metric shows the topic popularity in the development domain at monthly level and overall.
General popularity of topic	The metric provides information on the general interest in the topic (search popularity of the particular topic) on a monthly level and overall.
DataBench index per tool/technology	The metric illustrates the popularity and visibility of the specific tool or technology on monthly level and overall. The primary metric objective is to provide the user with highly important and emerging tools and technology based on combined research and development, industry and general interest data.
Number of research papers per tool/technology	The metric shows the highly popular and emerging tools and technologies in the research domain on a monthly level and overall.
Number of jobs per tool/technology	The metric displays tools and technologies that are important for industry/production, as well as for job seekers. The metric can be viewed on a monthly level and overall.
Number of Cordis EU projects per tool/technology	The metric provides highly popular and emerging tools for research and development projects on a monthly basis and overall.
Number of Github projects per tool/technology	The metric displays highly popular and emerging tools for the development, on a monthly basis and overall.
General popularity of tool/technology	The metric presents the popularity of tools or technologies based on search results.

Table 5 DataBench Observatory Metrics

The metrics are implemented in the DataBench Observatory tool and can be interactively tracked. Below in the document we discuss the specific metric values and show trend snapshots from the prototype implementation.

4.2. DataBench Popularity Index

The **DataBench Popularity Index** is a measure for ranking topic terms in the area of Artificial Intelligence and Big Data (with a specific focus on tools and technologies related to Benchmarking). The DataBench Index is composed based on the following inputs:

- **Research component**, such as **papers** from the **Microsoft Academic Graph (MAG)** [12]; The number of mentions of the topic term in the papers from the Microsoft

Academic Graph. Within this component, the MAG taxonomy is used, what allows for categorization of individual topic terms.

- **Industry component**, such as **job advertisements** from **Adzuna** service [13]; The number of mentions of the individual topic terms in the job advertisements.
- **Research and Development component**, such as **Horizon2020/FP7** projects from EU [14]; The number of mentions of individual topic terms in the descriptions of Horizon 2020 projects.
- **Media and Social media component**, such as **news** and **tweets** [15]; The number of mentions of the individual topic terms in news and tweets.
- **Technical Development component**, such as **projects** on **Github** [16]; The number of mentions of the individual topic terms in technical descriptions of the projects of Github.
- **General Interest**, such as **Google Trends** [17]; The frequency of searches of the individual topic term in the Google Trends.
- The additional dataset can include financial/investment data, such as **Preqin** [11].

The combined value is calculated with normalization and averaging of different component values and represents the relative value for term comparison within the ranking list.

Figure 14: DataBench Index Components presets the graphical description of the DataBench Popularity Index.

4.3. Methodology Pipeline

The DataBench methodology pipeline includes a number of steps for aggregating different data sources, calculating the DataBench Index and displaying the results:

1. Expanding Microsoft Academic Graph (MAG) taxonomy with DataBench terms. In this step generalization of the DataBench Ontology of topics related to AI from MAG topics is obtained, as well as extended with topic terms specifically defined for benchmarking.
2. Semantic annotation of textual data sources. Wikification is used for semantic annotation of textual data sources, such as job descriptions, news, project descriptions, technical notes.

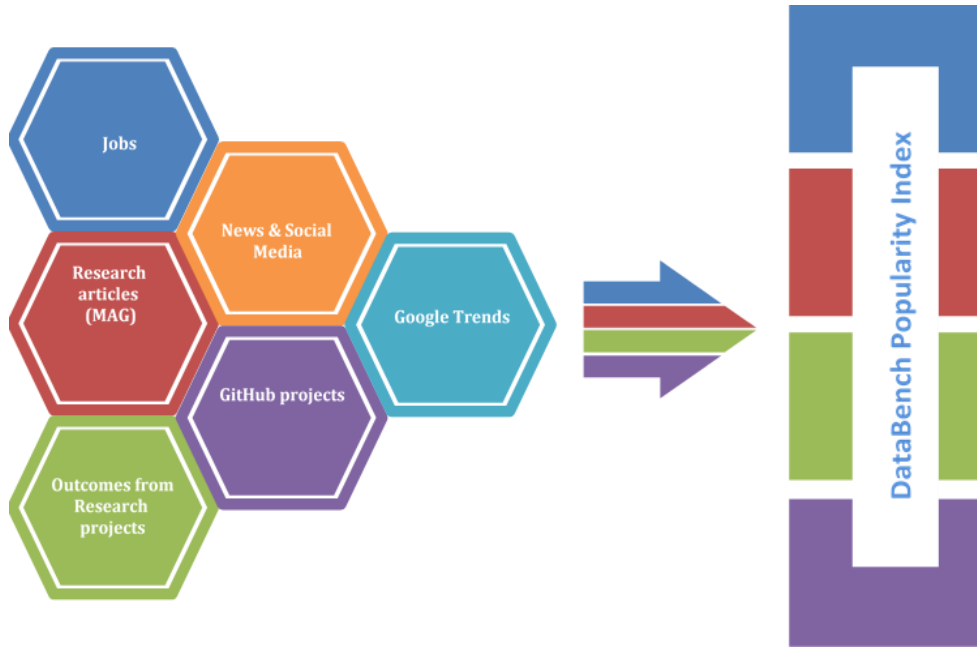


Figure 14: DataBench Index Components

3. Annotating data sources with topic terms from DataBench Ontology. In particular, MAG research papers (that are already pre-annotated with MAG taxonomy) are additionally annotated with expanded topic terms. Wikified data are used for mapping to DataBench Ontology for other resources.
4. Data normalization. Normalization is used for different components in order to produce the DataBench Index.
5. Data aggregation and generation of the DataBench Index.
6. Producing additional functionalities for different components:
 - a. MAG time series per term;
 - b. EU (Horizon 2020/FP7) projects time series per term;
 - c. News time series per term;
 - d. Google trends per term;
 - e. Job postings per term;
 - f. GitHub time series per term.

Figure 15: DataBench Methodology Aspects highlights a number of DataBench methodology main aspects. While Figure 16: Pipeline Challenges demonstrates the pipeline challenges, that should be accounted for.

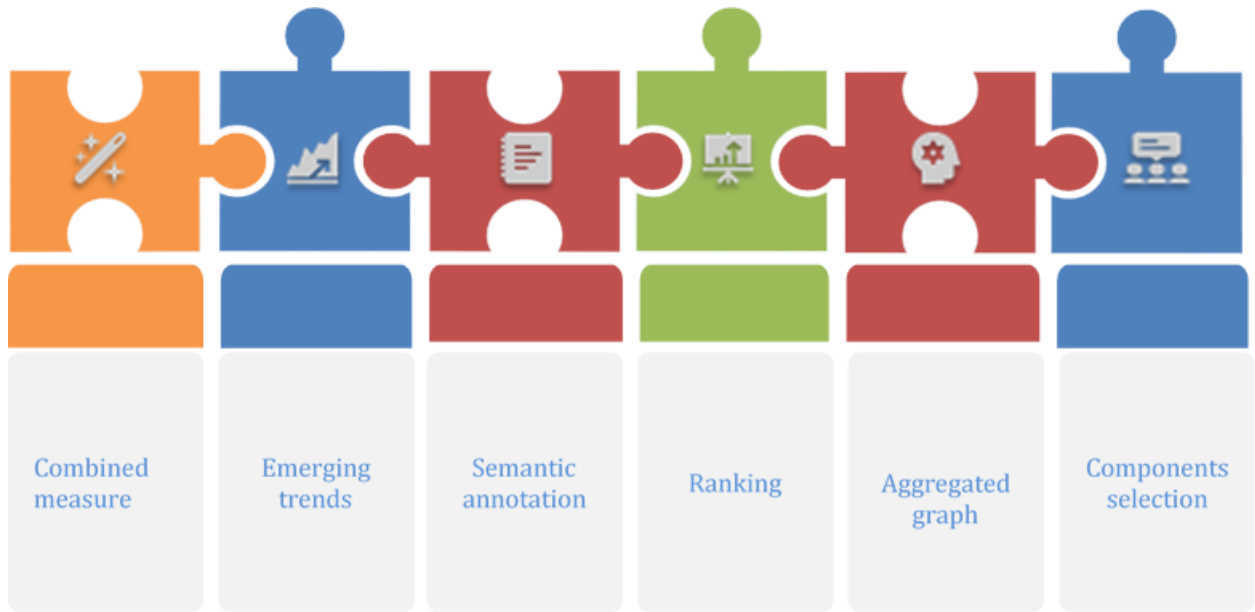


Figure 15: DataBench Methodology Aspects

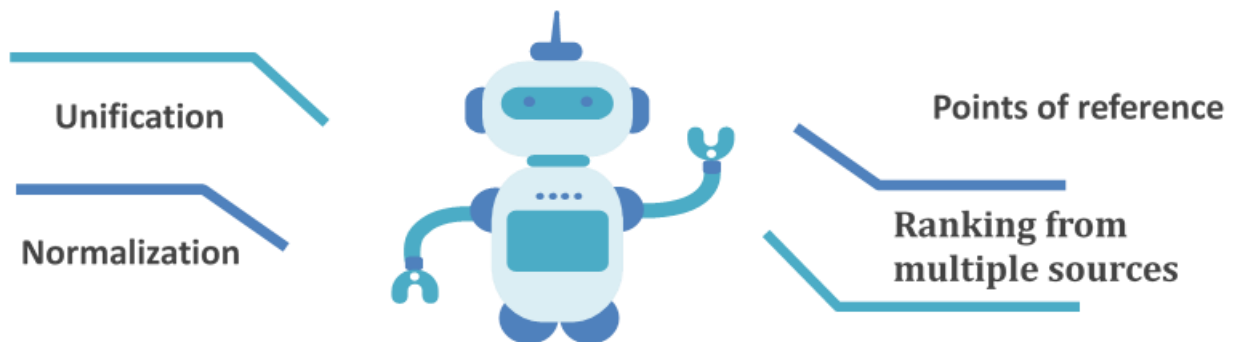


Figure 16: Pipeline Challenges

Below we describe the data sources with examples, as well as provide specific methodology and development elements.

4.4. Data Sources

The Data Sources used for the DataBench Observatory cover research and development, industry and general interest domains and include:

- Research papers from Microsoft Academic Graph;
- News and Social media from Event Registry system;
- Job postings from Adzuna service;
- European projects from CORDIS dataset (in particular, Horizon2020 EU projects);
- Projects (GitHub);
- Search trends from Google Trends.

The methodology and developed prototype are flexible for adding new data sources. For instance, in the future work we consider augmenting the data sources list with financial/investment data available from Preqin service [11].

Next we describe each data source in details and provide example of the original data.

Microsoft Academic Graph

The Microsoft Academic Graph (MAG) is a heterogeneous graph containing scientific publication records, citation relationships between those publications, as well as authors, institutions, journals, conferences, and fields of study.

For the purposes of DataBench project we have used data in English language, from all countries around the world.

MAG data ranges from year 1980. In DataBench project data under computer science category from April 2017 to present time was used.

In particular, 56m items have been examined, with 1.5m+ tagged with at least one of the tools since April 2017. Below an example MAG entry used for analysis can be viewed:

```
{
  "ID": "2607835903",
  "docType": "conference",
  "title": "a deep learning framework using passive wifi sensing for respiration monitoring",
  "abstract": "This paper presents an end-to-end deep learning framework using passive WiFi sensing to classify and estimate human respiration activity. A passive radar test-bed is used with two channels where the first channel provides the reference WiFi signal, whereas the other channel provides a surveillance signal that contains reflections from the human target. Adaptive filtering is performed to make the surveillance signal source-data invariant by eliminating the echoes of the direct transmitted signal. We propose a novel convolutional neural network to classify the complex time series data and determine if it corresponds to a breathing activity, followed by a random forest estimator to determine breathing rate. We collect an extensive dataset to train the learning models and develop reference benchmarks for the future studies in the field. Based on the results, we conclude that deep learning techniques coupled with passive radars offer great potential for end-to-end human activity recognition.",
  "date": "2017-12-01",
  "publisher": "IEEE",
  "citation": 14
}
```

Job advertisements

Job advertisement data come from Adzuna service, which is an employment website for job postings. Adzuna covers such countries, as Austria, Australia, Brazil, Canada, Germany, France, India, Italy, Netherland, New Zealand, Poland, Russia, Singapore, United Kingdom, USA, South Africa.

The job ads data are multi-lingual and are posted in English, German, Portuguese, French, Hindi, Italian, Dutch, Polish, Russian language.

In DataBench project we are using data from April 2017 to present time.

Each job posting is tagged by Adzuna to one of the following categories:

Consultancy, Charity & Voluntary, Property, IT, Legal, Customer Services, Teaching, Other/General, Accounting & Finance, Retail, Manufacturing, Hospitality & Catering, Healthcare & Nursing, Trade & Construction, Domestic help & Cleaning, Creative & Design, Logistics & Warehouse, HR & Recruitment, PR, Advertising & Marketing, Social work, Travel, Energy, Oil & Gas, Maintenance, Scientific & QA, Graduate, Engineering, Part time, Unknown, Sales, Administration.

For DataBench purposes, we selected all job postings from the IT category, with ~3.6m jobs tagged with at least one tool/topic.

Below and example of Adzuna job posting can be viewed:

```
{
  "id": 1590424001,
  "cat_id": 2,
  "loc_id": 191269,
  "date": "2020-07-01 19:28:03",
  "text": "Web UI Software Developer. WEB UI SOFTWARE DEVELOPER RESPONSIBILITIES Create web interfaces, using standard HTML/CSS practices, incorporating data from various Back End databases and distributed services. Create well-designed, tested code using best practices for website development, including mobile. Interact with stakeholders to work quickly and effectively to complete small edits requested by users, develop plans for larger projects and suggest new solutions to improve existing websites. Develop and maintain Back Office services including: batch processing, clearing, allocations; interacting with various 3rd party services. QUALIFICATIONS Bachelor's degree in a technical field or equivalent work experience Experience with web UI libraries and charting libraries(Bootstrap, D3, etc.) Working knowledge of web Servers like Apache. Working knowledge of Dockers and Containers. Strong Python programming skills. Comfortable with UNIX environment/basic Unix commands. Basic database knowledge (Sybase, MySQL, InfluxDB). Solid experience with PHP, HTML, CSS, Javascript. Ability to work under pressure within a dynamic trading environment. Attention to detail for problem solving and code robustness. If this is an opportunity that you're interested in please email your resume to: (see below)",
  "salary": "USD 80000.00 to 125000.00 per annum",
  "curr": "USD",
  "company": "Request Technology Kyle Honn",
  "country": "US",
  "Language": "en",
  "wikifierConcepts": [
    {"concept": "Unix", "pageRank": 0.1378616407144},
    {"concept": "MySQL", "pageRank": 0.125972536688128},
    {"concept": "HTML", "pageRank": 0.12535407381015098},
    {"concept": "InfluxDB", "pageRank": 0.107068096694765},
    {"concept": "PHP", "pageRank": 0.10671901410967},
    {"concept": "Python (programming language)", "pageRank": 0.0933367462579915},
    {"concept": "Batch processing", "pageRank": 0.0871312901153647},
    {"concept": "Bachelor's degree", "pageRank": 0.0866039527894413},
    {"concept": "Sybase", "pageRank": 0.0682187342471607},
  ]
}
```

```

    {"concept": "Cascading Style Sheets","pageRank": 0.0617339145729283
  }
]
};

```

News

In DataBench project we are using news data from Event Registry system. Event Registry is the world's leading news intelligence platform, empowering organizations to keep track of world events and analyze their impact.

The news data are cross-lingual and come in a variety of languages, such as English, German, Spanish, Catalan, Portuguese, Italian, French, Russian, Arabic, Turkish, Chinese, Japanese, Slovene, Croatian, Serbian, Bosnian, Albanian, Macedonian, Czech, Slovak, Polish, Basque, Hungarian, Dutch, Swedish, Finnish, Norwegian, Irish, Danish, Greek, Maltese, Romanian, Bulgarian, Lithuanian, Latvian, Estonian, Ukrainian, Armenian, Georgian, Azerbaijani, Persian, Indonesian, Thai, Hindi, Urdu.

In particular, while analyzing the occurrence of tools or technologies in news, we have established that out of ~1700 tools, around 600 tools have Wikipedia concepts that existed in Event Registry database. Using wikification process and DataBench ontology allows for identification of all tools and technologies from news.

The concept search allowed for searching in all the languages supported. Whereas for the rest, we used the tool term as a keyword for search, which limited the results to Latin languages only.

For each month, a query has been constructed for each tool to count the number of news articles that contain the concept/keyword and have one of the following concepts as well: artificial intelligence, database, company, or software. The reason behind that is to filter the results for tools that have a name with different meaning depending on the field. Like graphene, snowflakes, etc.

For topics, the results were aggregated from the results of the tools that fall under the topic. Below an Event Registry result can be observed:

```

{"articles":
{"page":1,
"totalResults":2866,
"pages":2866,
"results":
[ {"uri":"6305481961",
"lang":"eng",
"isDuplicate":false,
"date":"2020-11-19",
"time":"09:15:00",

```

```

"dateTime":"2020-11-19T09:15:00Z",
"dateTimePub":"2020-11-19T09:15:00Z",
"dataType":"news",
"sim":0,
"url":"https://npinvestor.dk/node/443493",
"title":"GridGain Systems Named to Deloitte 2020 Technology Fast 500™ for Third Consecutive Year",
"body":"GridGain Ranked 380th Fastest Growing Company Nationally with 253 Percent Revenue Growth Over the Past Four Years, Ranked 73rd in the San Francisco Bay Area\n\nFOSTER CITY, Calif., Nov. 19, 2020 (GLOBE NEWSWIRE)...",
"source":
{
  "uri":"npinvestor.dk",
  "dataType":"news",
  "title":"npinvestor.dk"},
"authors":[],
"image":null,
"eventUri":null,
"sentiment":0.4823529411764707,
"wgt":343473300,
"relevance":52
}}}
```

Cordis EU projects

CORDIS dataset covers EU research project from Horizon2020 and FP7 programs. In total we have obtained ~55k projects, ~51k of which are tagged with at least one tool/topic.

For each project, the description abstract was wikified and the main concepts were extracted, the concepts were matched with the DataBench ontology concepts of the tools and topics. Since the projects last for a long span (typically 3 years), the counts are normalized per month, i.e. if a project lasts over 3 years, 1/36 score is added to each of its 36 months. Below we provide an example of how project data can be viewed:

```

{"projectUrl":"","
"coordinator":"AARHUS UNIVERSITET",
"acronym":"BTVI",
"endDate":"2019-03-31",
"topics":"ERC-CG-2013-PE8",
"subjects":[],
"coordinatorCountry":"DK",
"title":"First Biodegradable Biocatalytic VascularTherapeutic Implants",
"objective":"We aim to perform academic development of a novel biomedical opportunity: localized synthesis of drugs within biocatalytic therapeutic vascular implants (BVI) for site-specific drug
```


delivery to target organs and tissues. Primary envisioned targets for therapeutic intervention using BVI are atherosclerosis, viral hepatitis, and hepatocellular carcinoma: three of the most prevalent and debilitating conditions which affect hundreds of millions worldwide and which continue to increase in their importance in the era of increasingly aging population. For hepatic applications, we aim to develop drug eluting beads which are equipped with tools of enzyme-prodrug therapy (EPT) and are administered to the liver via trans-arterial catheter embolization. Therein, the beads perform localized synthesis of drugs and imaging reagents for anticancer combination therapy and theranostics, antiviral and anti-inflammatory agents for the treatment of hepatitis. Further, we conceive vascular therapeutic inserts (VTI) as a novel type of implantable biomaterials for treatment of atherosclerosis and re-endothelialization of vascular stents and grafts. Using EPT, inserts will tame the guardian of cardiovascular grafts, nitric oxide, for which localized, site specific synthesis and delivery spell success of therapeutic intervention and/or aided tissue regeneration. This proposal is positioned on the forefront of biomedical engineering and its success requires excellence in polymer chemistry, materials design, medicinal chemistry, and translational medicine. Each part of this proposal - design of novel types of vascular implants, engineering novel biomaterials, developing innovative fabrication and characterization techniques is of high value for fundamental biomedical sciences. The project is target-oriented and once successful, will be of highest practical value and contribute to increased quality of life of millions of people worldwide."

```
"call":"ERC-2013-CoG",
"participantCountries":[],
"fundingScheme":"ERC-CG",
"ecMaxContribution":"1996126",
"id":"617336",
"rcn":"185654",
"programme":"FP7-IDEAS-ERC",
"frameworkProgramme":"FP7",
"startDate":"2014-04-01",
"totalCost":"1996126",
"status":"ONG",
"participants":[]}
```

Google Trends

In DataBench we are using trending score from Google trends for the period April 2017 till now. The score shows how the search trend for the term is going up and down in relation to the selected period. The peak point in time for each term is given a score of 100, and the rest of the scores are normalized with respect to that peak point. Below a Google trend example can be viewed:

```
{"ID": "databench.ijs.si/apache_spark",
"term": "Apache Spark",
"short_term_graph":
{"lines":
[{"term": "Apache Spark",
"points":
[{"value": 98, "date": "2017-04-02"},
{"value": 94, "date": "2017-04-09"},
```

```
{
  "value": 83, "date": "2017-04-16"},
  {"value": 95, "date": "2017-04-23"},
  {"value": 92, "date": "2017-04-30"},
  {"value": 81, "date": "2017-05-07"},
  {"value": 95, "date": "2017-05-14"},
  {"value": 85, "date": "2017-05-21"},
  {"value": 97, "date": "2017-05-28"},
  {"value": 94, "date": "2017-06-04"},
  {"value": 92, "date": "2017-06-11"},...]
}]}
```

GitHub

GitHub is a development platform that allows for hosting open source as well as business projects. The users can host and review code, manage projects, and build software alongside 50 million developers. In DataBench project the source of data is the list of repositories related to AI/ML, where the repository has at least 5 stars or forks.

In the current prototype the total number of such repositories is 14k (with a possibility for extension in the future). We tagged a repository with a tool or topic if it is mentioned in any of the textual description files (README, etc.), with total of ~12k tagged with at least one tool/topic. The range of data used is from April 2017 till now.

Below a GitHub example can be observed:

```
{
  "repository": "tensorflow/tensorflow",
  "contents":
    [
      {
        "name": "README.md",
        "path": "README.md",
        "sha": "63d85ce2df4a9ae0a7303a627f28561410d0ddf1",
        "size": 19859,
        "url": "https://api.github.com/repos/tensorflow/tensorflow/contents/README.md?ref=master",
        "download_url": "https://raw.githubusercontent.com/tensorflow/tensorflow/master/README.md"
      },
      {
        "name": "RELEASE.md",
        "path": "RELEASE.md",
        "sha": "d73bf4ca0462e91839169074cdf0cde0485e333c",
        "size": 316026,
        "url": "https://api.github.com/repos/tensorflow/tensorflow/contents/RELEASE.md?ref=master",
        "download_url": "https://raw.githubusercontent.com/tensorflow/tensorflow/master/RELEASE.md"
      }
    ]
  "file_contents":
    {
      "README.md":
        "[TensorFlow](https://www.tensorflow.org/) is an end-to-end open source platform\nfor machine learning. It has a comprehensive, flexible ecosystem"
    }
  }
```

```
of\n[tools](https://www.tensorflow.org/resources/tools),\n[libraries](https://www.tensorflow.org/resources/libraries-extensions), and\n[community](https://www.tensorflow.org/community) resources that lets\nresearchers push the state-of-the-art in ML and developers easily build and\nndeploy ML-powered applications.\n\nTensorFlow was originally developed by researchers and engineers working on the\nGoogle Brain team within Google's Machine Intelligence Research organization to\nconduct machine learning and deep neural networks research. The system is\ngeneral enough to be applicable in a wide variety of other domains, as well.\n\nTensorFlow provides stable [Python](https://www.tensorflow.org/api_docs/python)\nand [C++](https://www.tensorflow.org/api_docs/cc) APIs,...",...},\n"created at": "2015-11-07T01:19:20Z"}
```

4.5. Data Dimensions and Data Formats

In the process of developing the DataBench Observatory some of main data dimensions include time series for topic terms (related to Artificial Intelligence and Big Data), country and date.

Table 6 Data Sources **Format** presents CSV file per data source, each containing the following columns:

COLUMN	TERM	DATE	VALUE
Format	Topic Display Name	YYYY/MM	DOUBLE
Example	Artificial Intelligence	2020/08	5

Table 6 Data Sources Format

4.6. Pillars

In the DataBench Observatory we have defined two primary pillars for the purpose of defining topic term popularity, importance and visibility.

First of all, the observatory provides **ranking based on absolute numbers** (obtained via DataBench Index). In particular, there are two types of views, based on date dimension:

- overall
- by month

In addition, we have defined:

- ranking for topics (based on all available topics in the DataBench Ontology) and
- ranking for tools and technologies (represented as ontology leaves).

The time series/overall for ranking based on absolute numbers include:

- MAG number of papers time series;
- number of papers (normalized per affiliation) per term;
- news and social media/Event Registry time series;
- number of news articles per month per term;
- Adzuna time series;

- number of job posting per country per month per term;
- CORDIS time series;
- number of projects per month per term;
- GitHub time series;
- number of GitHub projects per month per term.

The second pillar displays the **ranking based on the emerging topics**. Within this pillar the growth ratio instead of absolute numbers, breakout concepts in each source, a ratio based on the previous period are calculated.

The time series for ranking based on the emerging topics additionally include:

- Google Trends time series;

4.7. DataBench Ontology and Knowledge Graph Monitoring

The **DataBench Aggregated Graph** is a knowledge graph built upon several aggregated resources applicable for DataBench project and described above as data sources.

DataBench ontology

In order to develop the DataBench Aggregated Graph we have composed an **ontology** based on Artificial Intelligence, Big Data, Benchmarking related topics from Microsoft Academic Graph and extended/populated the ontology with tools and technologies from the relevant areas.

Microsoft Academic Graph (MAG) taxonomy has been expanded with DataBench terms – over 1,700 tools and technologies related to Benchmarking, Big Data, Artificial Intelligence.

New concepts have been aligned with MAG topic, MAG keyword, Wikipedia (for analysis in wikification) and Event Registry concepts.

Below we provide an example of aligning DataBench ontology with Wiki concepts and ER concepts.

Term: Apache OPEN NLP
Parent: [databench.ijs.si/nlp]
URI: databench.ijs.si/apache_open_nlp
Wiki concept: https://en.wikipedia.org/wiki/Apache_OpenNLP
ER concept: “Apache OPEN NLP”

The DataBench ontology is used in the semantic annotation process of the unstructured textual information from the available data sources.

Wikification or semantic annotation with Wikipedia concepts is an intermedia step in the process of annotation with DataBench ontology.

In wikification a global disambiguation method based on constructing a mention-concept graph and computing PageRank over it is used to identify a coherent set of relevant concepts for the document. Since several of our data sources are multi-lingual, wikification is a

suitable approach that supports any language for which a sufficiently large Wikipedia is available.

The task of wikifying [18] an input document can be broken down into several closely interrelated subtasks: (1) identify phrases (or words) in the input document that refer to a Wikipedia concept; (2) determine which concept exactly a phrase refers to; (3) determine which concepts are relevant enough to the document as a whole that they should be included in the output of the system (i.e. presented to the user).

As a result of wikification, each item from the dataset (with available textual information) obtains a set of possible annotations from Wikipedia. Since DataBench ontology is aligned with Wikipedia and Event Registry concepts, following the wikification process, we automatically obtain a semantically annotated documents with DataBench ontology.

Similar approach applies to the news dataset (from Event Registry), pre-annotated with Event Registry concepts. Since DataBench ontology is aligned with ER concepts, we automatically obtain news annotations with DataBench ontology.

DataBench Aggregated Graph

In DataBench Deliverable 5.1 Initial Evaluation of DataBench Metrics [4] we have discussed an option of developing the knowledge graph representation of DataBench data.

The modelled ontology contributes to the development of the DataBench Aggregated Graph that aligns the knowledge obtained from different available data sources.

In particular, the entities of the graph represent tools, topics, Github repositories, news articles, job postings, EU research projects, Categories et al.

Annex: DataBench Ontology Formalization presents an overview of the Aggregated Graph development. In our work on the aggregated graph we have focused on representing topics and tools (from the DataBench Observatory), providing the appropriate information about topic/tools and technology popularity.

The DataBench ontology and Aggregated Graph are published in Github on the following link:

<https://github.com/besher-massri/DataBenchKnowledgeGraph>

The visual representation of the DataBench ontology can be viewed on:

https://github.com/besher-massri/DataBenchKnowledgeGraph/blob/main/databench_ontology.svg

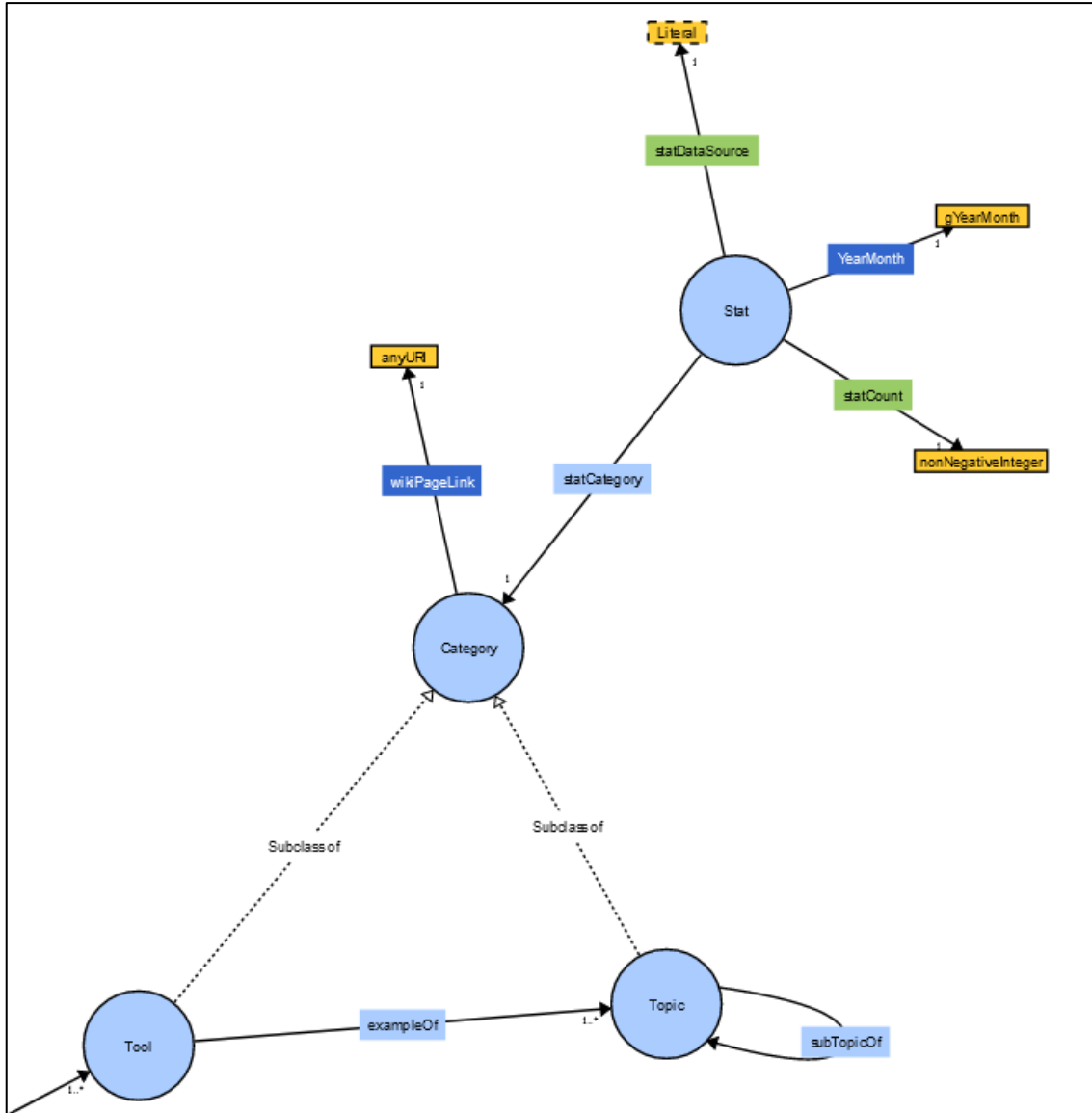


Figure 17: DataBench Ontology Snapshot

Figure 17: DataBench Ontology Snapshot presents a snapshot from the DataBench ontology related to tools and topics.

Using DataBench ontology and DataBench Aggregated Graph provides a knowledge sharing option about topics, tools and technologies popularity and visibility in different data sources.

5. DataBench Observatory Implementation

5.1. Frontend and Backend

The **DataBench Observatory** implementation is based on Backend and Frontend components.

The Frontend component includes a number of elements, such as:

- Dashboard information
- Topics
 - Topics popularity index
 - Indicators
 - Topics timeseries
 - Data volume timeseries
- Tools
 - Tools popularity index
 - Tools timeseries
- BDVE
 - Tools ranking in big data projects
 - Tools timeseries in big data projects
 - Big data project details
- News
 - AI live news
- Data sources
- Methodology
- Knowledge graph
- About
- Contact

The functionalities of the observatory are set to:

- Provide terms ranking as table;
- Provide search bar for the term search;
- Provide possibility of choosing the sources included in the ranking.
- Gap minder visualization with options to choose the axes;
- Time series visualized as a line graph with options to select the source;
- Provide similar projects from big data projects collection;
- Live news.

The program architecture is a pipeline that consists of crawlers, analyzer, and service provider - the crawlers consists of python scripts for crawling the data from the designated APIs.

Crawlers have been constructed for getting Google trends data, Github data, and Event Registry data (news). Whereas for Adzuna (job posting) and MAG (research paper), a periodic dump of all the data is being acquired.

The analyzer parses the data and do the necessary processing to generate the needed statistics. The analyzer is implemented in C++ and R - the service provider is a web application that uses the outputted statistics from the analyzer and visualize it in the DataBench Observatory dashboard.

The DataBench Observatory dashboard is built in node.js for backend, with express.js for web server, and d3.js library for the visualizations on the front end.

5.2. DataBench Observatory Documentation

In this section we document the explanatory options for the interface of the DataBench Observatory. We provide information on the methodology and data sources used for the DataBench Popularity Index and define key visualizations for the observatory.

On the left side of the DataBench Observatory, demonstrated by Figure 18: **DataBench Observatory (Explanatory menu, accessed in May 2021)**, the users can find an explanatory menu with a number of pages. The Data Sources page describes in detail the available data sources. The Methodology page provides users with insights into the DataBench Observatory methodology, while the Knowledge Graph page demonstrates the semantic technologies used in the process of observatory development.

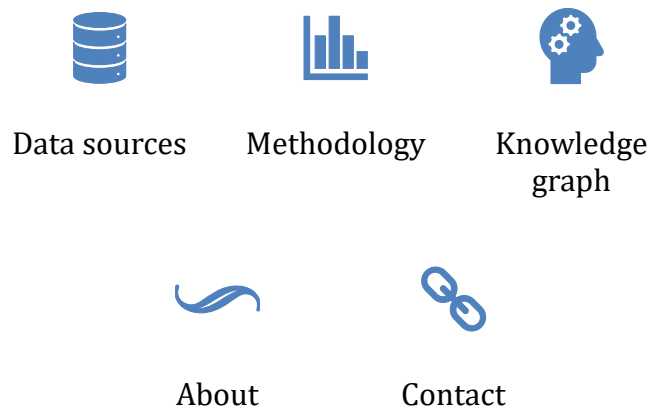


Figure 18: DataBench Observatory (Explanatory menu, accessed in May 2021)

The About page summarizes the information about DataBench Observatory developers and provides a link to the DataBench project website. Users can contact the developers through the contact details included in the Contact page.

Figure 19: **DataBench Observatory (Information page, accessed in May 2021)** provides a view on the dashboard information – introduction to the observatory, a list of available data sources, observatory sections and key visualizations (with detailed explanations of observatory usage), summary of the information.

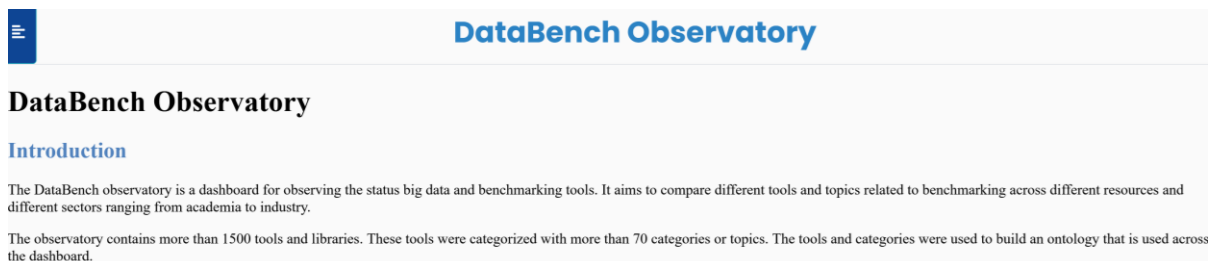


Figure 19: DataBench Observatory (Information page, accessed in May 2021)

Figure 20: **DataBench Observatory (Data sources, accessed in May 2021)** documents a detailed description of the data sources used in the development process. In particular, users can obtain the information about the timeline used for each data source, the update process and applied methods.

Data Sources

The Data Sources used for the DataBench Observatory cover research and development, industry and general interest domains and include:

- Research papers from Microsoft Academic Graph;
- News and Social media from Event Registry system;
- Job postings from Adzuna service;
- European projects from CORDIS dataset (in particular, Horizon2020 EU projects);
- Projects (GitHub);
- Search trends from Google Trends.

The methodology and developed prototype are flexible for adding new data sources. For instance, in the future work we consider augmenting the data sources list with financial/investment data available from Preqin service.

Next we describe each data source in details and provide example of the original data.

News and Social media from Event Registry system.

Job postings from Adzuna service.

European projects from CORDIS dataset (in particular, Horizon2020 EU projects).

Big Data Projects: a selected list of Big data projects that is a subset of CORDIS.

Open source projects from GitHub.

Search trends from Google Trends.

Figure 20: DataBench Observatory (Data sources, accessed in May 2021)

Users can access the Knowledge Graph page including the description of the DataBench ontology, the visualization of the Knowledge Graph and the methodologies related to the applied semantic technologies (Figure 21: **DataBench Observatory (Ontology description, accessed in May 2021)**).

DataBench Ontology and Knowledge Graph Monitoring

The DataBench Aggregated Graph is a knowledge graph built upon several aggregated resources applicable for DataBench project and described above as data sources.

DataBench Ontology

In order to develop the DataBench Aggregated Graph we have composed an ontology based on Artificial Intelligence, Big Data, Benchmarking related topics from Microsoft Academic Graph and extended/populated the ontology with tools and technologies from the relevant areas.

Microsoft Academic Graph (MAG) taxonomy has been expanded with DataBench terms – over 1,700 tools and technologies related to Benchmarking, Big Data, Artificial Intelligence.

New concepts have been aligned with MAG topic, MAG keyword, Wikipedia (for analysis in wikification) and Event Registry concepts.

The DataBench ontology is used in the semantic annotation process of the unstructured textual information from the available data sources. Wikification or semantic annotation with Wikipedia concepts is an intermediate step in the process of annotation with DataBench ontology.

In wikification a global disambiguation method based on constructing a mention-concept graph and computing PageRank over it is used to identify a coherent set of relevant concepts for the document. Since several of our data sources are multi-lingual, wikification is a suitable approach that supports any language for which a sufficiently large Wikipedia is available.

The task of wikifying an input document can be broken down into several closely interrelated subtasks:

1. identify phrases (or words) in the input document that refer to a Wikipedia concept;
2. determine which concept exactly a phrase refers to;
3. determine which concepts are relevant enough to the document as a whole that they should be included in the output of the system (i.e. presented to the user).

As a result of wikification, each item from the dataset (with available textual information) obtains a set of possible annotations from Wikipedia. Since DataBench ontology is aligned with Wikipedia and Event Registry concepts, following the wikification process, we automatically obtain a semantically annotated documents with DataBench ontology.

Figure 21: DataBench Observatory (Ontology description, accessed in May 2021)

The methodology about the DataBench Popularity Index and pipeline are documented in detail on the Methodology page (Figure 22: **DataBench Observatory (Methodology description, accessed in May 2021)**).

Methodology

The DataBench Popularity Index

The DataBench Popularity Index is a measure for ranking topic terms in the area of Artificial Intelligence and Big Data (with a specific focus on tools and technologies related to Benchmarking). The DataBench Index is composed based on the following inputs:

- Research component, such as papers from the Microsoft Academic Graph (MAG); The number of mentions of the topic term in the papers from the Microsoft Academic Graph. Within this component, the MAG taxonomy is used, what allows for categorization of individual topic terms.
- Industry component, such as job advertisements from Adzuna service; The number of mentions of the individual topic terms in the job advertisements.
- Research and Development component, such as Horizon2020/FP7 projects from EU; The number of mentions of individual topic terms in the descriptions of Horizon 2020 projects.
- Media and Social media component, such as news and tweets; The number of mentions of the individual topic terms in news and tweets.
- Technical Development component, such as projects on Github; The number of mentions of the individual topic terms in technical descriptions of the projects of Github.
- General Interest, such as Google Trends; The frequency of searches of the individual topic term in the Google Trends.
- The additional dataset can include financial/investment data, such as Preqin.

The combined value is calculated with normalization and averaging of different component values and represents the relative value for term comparison within the ranking list.

The DataBench methodology pipeline

The DataBench methodology pipeline includes a number of steps for aggregating different data sources, calculating the DataBench Index and displaying the results:

1. Expanding Microsoft Academic Graph (MAG) taxonomy with DataBench terms. In this step generalization of the DataBench Ontology of topics related to AI from MAG topics is obtained, as well as extended with topic terms specifically defined for benchmarking.
2. Semantic annotation of textual data sources. Wikification is used for semantic annotation of textual data sources, such as job descriptions, news, project descriptions, technical notes.
3. Annotating data sources with topic terms from DataBench Ontology. In particular, MAG research papers (that are already pre-annotated with MAG taxonomy) are additionally annotated with expanded topic terms. Wikified data are used for mapping to DataBench Ontology for other resources.

Figure 22: DataBench Observatory (Methodology description, accessed in May 2021)

The Data Volume Timeseries page provides details on data distributions over time for different data sources (Figure 23: **DataBench Observatory (Data distributions, accessed in May 2021)**). The users can pick up a data source (from Resources), as well as select Scale and Cumulative options for the graph representation.

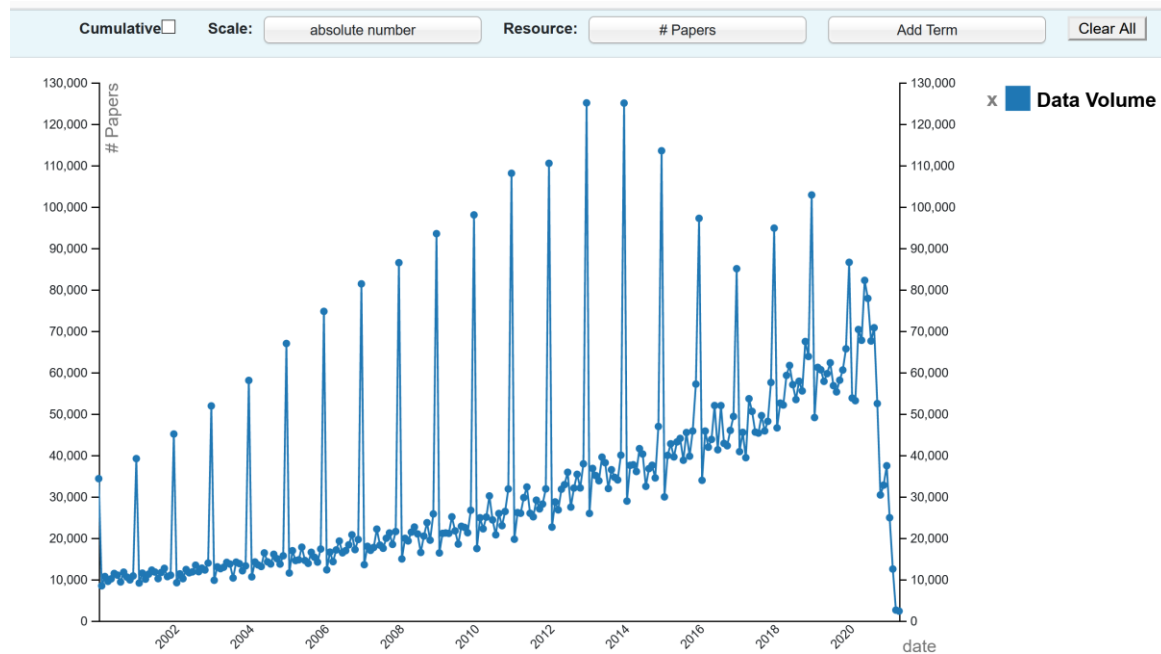


Figure 23: DataBench Observatory (Data distributions, accessed in May 2021)

Additionally, as stated above, the DataBench Observatory tool currently provides users with a list of key visualizations (with explanations for each visualization). Each visualization is documented – users can easily get insights into the tool usage. In particular, users can get a distribution time series for each topic or tool over time, check the live news or view the tools and topics popularity in different data sources.

5.3. DataBench Observatory Usage Scenarios and Metrics Results

In order to illustrate the functionalities of DataBench Observatory, we have developed possible usage scenarios for researcher, industry and general public.

DataBench Observatory for Academic users

With the DataBench Observatory the academic users/researchers have a possibility to explore popular and visible topics in the area of Artificial Intelligence, Big Data and Benchmarking. Since within the Observatory we analyze Microsoft Academic Graph content with millions of academic papers, the researchers have an option to map their research interests on the observatory landscape.

Researchers who participate in the EU projects have a possibility to observe popular topics within EU research and development domain, as well as topic trending in time.

Figure 24: **DataBench Popularity Index** shows the illustration of DataBench Index for topics (score 10 is the maximum normalized popularity).

month: All						
Search: <input type="text"/>						
Topic	Papers	EU Projects	News	Github	Jobs	Total
infrastructure	3.58	2.47	10	2.53	10	5.72
Artificial Intelligence	10	1.07	4.67	10	1.17	5.38
Machine Learning	8.35	1.07	4.3	8.67	1.17	4.71
customer	4.87	10	1.71	1.06	1	3.73
stat, tool	3.21	2.03	1.13	8.99	1.45	3.36
Database	2.07	1.44	1.44	1.6	8.9	3.09
label	7.72	1	3.84	1.54	1.01	3.02
relational dbms	1.57	1.07	1.21	1.26	8.25	2.67
misc	4.3	2.22	3.52	2.19	1.14	2.67
Analytics	4.17	2.72	1.38	1.63	1.69	2.32
search engines	4.97	2.58	1.28	1.16	1.46	2.29

Figure 24: DataBench Popularity Index (Topics, accessed in November 2020)

Figure 25: **Time Series (Topics, accessed in November 2020)** shows time series for topics from the areas of Artificial Intelligence, Big Data and Benchmarking.

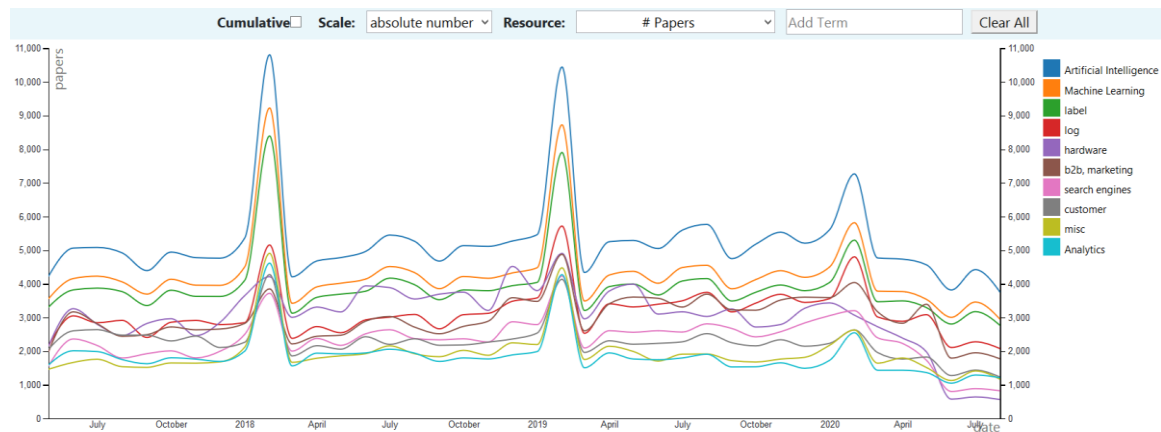


Figure 25: Time Series (Topics, accessed in November 2020)

For instance, researchers interested in “Artificial Intelligence” topic can observe its high popularity (score 10 is the maximum normalized popularity) within academic Papers.

DataBench Observatory for Industrial users

The industrial users at the same time, are interested in the tools and technologies available in the market. The DataBench Observatory gives an insight into the variety of popular tools and technologies and can provide a company with an overview of which tools and technologies are interesting for investment (in particular, which tools can be used in the production process, which tools are used by other companies from the same industry/category, what are the tools and technologies with emerging trends in coding repositories).

Figure 26: DataBench Popularity Index (November 2020) presents the DataBench Popularity Index for tools and technologies.

Topic	Categories	Papers	EU Projects	News	Github	Jobs	Search Volume	Total
Microsoft	infrastructure,computer_s cience	1.72	1.13	5.24	1.3	10	6.34	4.29
Google	infrastructure,computer_s cience	3.74	2.23	8.21	2.76	1.27	7.38	4.26
Amplitude	customer,computer_sci ence	5	10	1.03	1.04	1	7.11	4.2
Vector	log,computer_sci ence	10	1	1.19	2.32	1	9.35	4.14
Python	stat_tool,computer_sci ence	1.81	1.06	1.2	10	1	8.96	4.01
Node	b2b_marketing,computer _science	8.71	1	1.17	1.62	1	10	3.92
ARM	hardware,computer_sci ence	10	1	1.03	1.13	1	9.37	3.92

Figure 26: DataBench Popularity Index (November 2020)

Figure 27: Time Series (Tools and Technologies, accessed in November 2020) shows time series for tools and technologies from the areas of Artificial Intelligence, Big Data and Benchmarking.

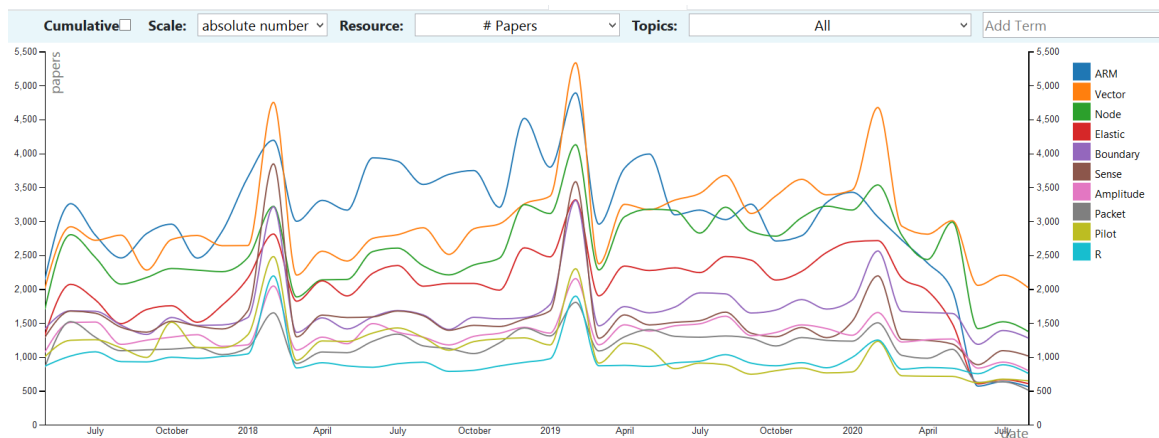


Figure 27: Time Series (Tools and Technologies, accessed in November 2020)

For instance, based on the figures above it is possible to notice that Microsoft tools are highly requested in job postings, Python is the most popular language at GitHub and users on the web search a lot for Node.js solutions.

DataBench Observatory for General public

The DataBench Observatory provides a possibility of exploring topic evolution with respect to different dimensions. The users can choose the dimensions from available data sources and view the changing topics.

Figure 28: Topic Evolution Visualization (Indicators) displays an example of the topic evolution.

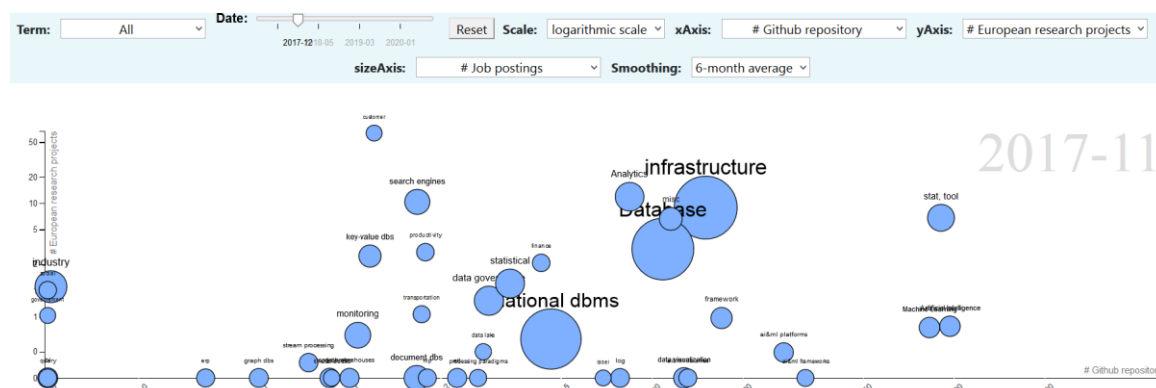


Figure 28: Topic Evolution Visualization (Indicators, accessed in November 2020)

Another functionality interesting for the general public users is the possibility to dynamically observe news about relevant topics, tools and technologies.

Figure 29: Live News presents a stream of live news related to Artificial Intelligence.

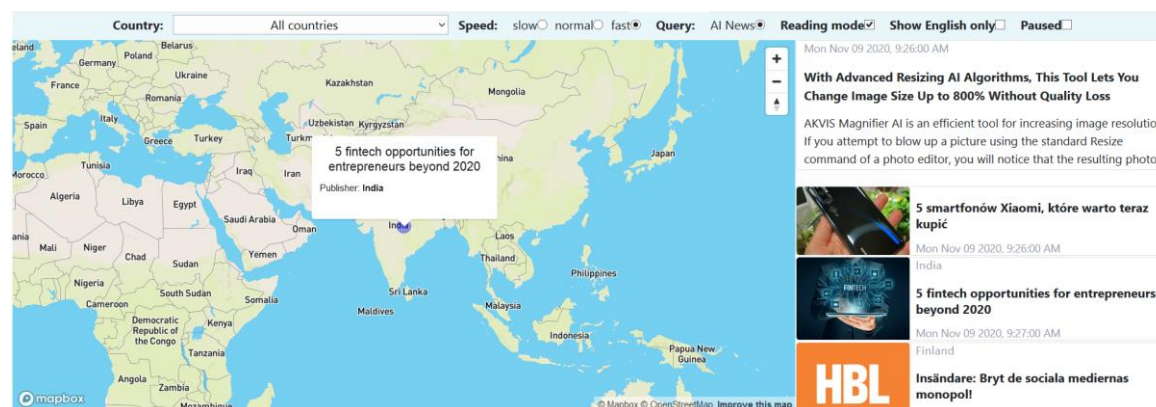


Figure 29: Live News (accessed in November 2020)

DataBench Observatory metrics results

For illustration, Table 7 DataBench Observatory Metrics Results demonstrates metric values for a number of popular topics, tools and technologies in the area of Artificial Intelligence, Big Data and Benchmarking.

The DataBench Observatory provides a search functionality for each topic, tool or technology from the ontology. The user has the possibility to obtain metric value via search, as well as to draw trending graphs for topics, tools and technologies.

TOPIC/TOOL/TECHNOLOGY	METRIC	METRIC VALUES	
		per month (07.2020)	overall
Machine Learning	DataBench index per topic	4.53	4.71
Machine Learning	Number of research papers per topic	8.07	8.35
Machine Learning	Number of jobs per topic	1.15	1.17
Machine Learning	Number of Cordis EU projects per topic	1.06	1.07
Machine Learning	Number of GitHub projects per topic	8.56	8.67
Machine Learning	Number of news per topic	3.81	4.3
Database	DataBench index per topic	3.12	3.09
Database	Number of research papers per topic	1.75	2.07
Database	Number of jobs per topic	9.36	8.9
Database	Number of Cordis EU projects per topic	1.42	1.44
Database	Number of GitHub projects per topic	1.53	1.6

Database	Number of news per topic	1.53	1.44
TensorFlow	DataBench index per tool/technology	6.14	6.02
TensorFlow	Number of research papers per tool/technology	4.68	3.17
TensorFlow	Number of jobs per tool/technology	10	10
TensorFlow	Number of Cordis EU projects per tool/technology	1	1
TensorFlow	Number of GitHub projects per tool/technology	10	10
TensorFlow	Number of news per tool/technology	3.48	2.65
TensorFlow	General popularity of tool/technology	7.66	9.27
Scala	DataBench index per tool/technology	1.25	1.77
Scala	Number of research papers per tool/technology	1.07	1.13
Scala	Number of jobs per tool/technology	1	1
Scala	Number of Cordis EU projects per tool/technology	1.03	1.04
Scala	Number of GitHub projects per tool/technology	1	1.08
Scala	Number of news per tool/technology	1.39	1.54

Scala	General popularity of tool/technology	2.03	4.83
HiBench	DataBench index per tool/technology	2.5	2.24
HiBench	Number of research papers per tool/technology	1	1.27
HiBench	Number of jobs per tool/technology	5.5	5.5
HiBench	Number of Cordis EU projects per tool/technology	1	1
HiBench	Number of GitHub projects per tool/technology	5.5	1
HiBench	Number of news per tool/technology	1	1
HiBench	General popularity of tool/technology	1	3.66

Table 7 DataBench Observatory Metrics Results

5.4. Integration into DataBench Toolbox

Figure 30: DataBench Observatory in DataBench Toolbox shows the location of the DataBench Observatory in the Toolbox.



Figure 30: DataBench Observatory in DataBench Toolbox

Summary

In DataBench Deliverable 5.2 we present a final set of DataBench indicators and introduce the DataBench Popularity Index – a quantitative measure for ranking the concepts and topics in the area of Artificial Intelligence and Big Data.

In particular, the document refers to the DataBench Ecosystem of Indicators that include both technical and business and industry relevant indicators, as well as quality metrics, and their implementation in the DataBench Toolbox.

Furthermore, we describe the functionalities of the DataBench Observatory - a tool for observing the popularity, importance and the visibility of topic terms.

The DataBench Popularity Index is built upon multiple data sources, including:

- Research papers from Microsoft Academic Graph;
- News and Social media from Event Registry system;
- Job postings from Adzuna service;
- European projects from CORDIS dataset (in particular, Horizon2020 EU projects);
- Projects (GitHub);
- Search trends from Google Trends.

The DataBench Observatory tool provides a user a possibility to observe a number of data-driven metrics, such as DataBench index per topic/tool/technology, Number of research papers per topic/tool/technology, Number of jobs per topic/tool/technology, Number of Cordis EU projects per topic/tool/technology, Number of GitHub projects per topic/tool/technology, General popularity of topic/tool/technology.

The functionalities of the DataBench Observatory allow users to view information on different pillars (ranking based on absolute numbers and ranking based on emerging topics).

In addition, we suggest the knowledge formalization possibilities and present the DataBench ontology for sharing knowledge about topics, tools and technologies popularity. The methodology behind the DataBench Popularity Index allows for automatic maintenance and regular updates based on the availability of data sources.

Bibliography

[1] DataBench deliverable D1.1 Industry Requirements with benchmark metrics and KPIs, <https://www.databench.eu/public-deliverables> (accessed in December 2020).

[2] DataBench deliverable D1.2 DataBench Framework - with Vertical Big Data Type benchmarks, <https://www.databench.eu/public-deliverables> (accessed in December 2020).

[3] DataBench deliverable D3.4 Release Version of DataBench Toolbox including Visualization and Search Components, <https://www.databench.eu/public-deliverables> (accessed in December 2020).

- [4] DataBench deliverable D5.1 Initial Evaluation of DataBench Metrics, <https://www.databench.eu/public-deliverables> (accessed in December 2020).
- [5] DataBench deliverable D5.3 Assessment of technical usability, relevance, scale and complexity, <https://www.databench.eu/public-deliverables> (accessed in December 2020).
- [6] DataBench deliverable D5.4 Analytic modelling relationships between metrics, data and project methodologies, <https://www.databench.eu/public-deliverables>.
- [7] DataBench deliverable D5.5 Final report on methodology for evaluation of industrial analytic projects scenarios, <https://www.databench.eu/public-deliverables>.
- [8] BDVA Reference Model, <https://databench.ijs.si/bdva> (accessed in December 2020).
- [9] MLPerf Training, <https://mlperf.org/training-overview> (accessed in December 2020).
- [10] MLPerf Inference, <https://mlperf.org/inference-overview> (accessed in December 2020).
- [11] Preqin - provider of data, analytics, and insights to the alternative assets community, <https://www.preqin.com> (accessed in December 2020).
- [12] Microsoft Academic Graph, <https://www.microsoft.com/en-us/research/project/microsoft-academic-graph> (accessed in December 2020).
- [13] Adzuna - employment website for job advertisements, <https://www.adzuna.co.uk> (accessed in December 2020).
- [14] Horizon2020 projects, <https://data.europa.eu/euodp/sl/data/dataset/cordisH2020projects> (accessed in December 2020).
- [15] Event Registry – news intelligence platform, <https://eventregistry.org> (accessed in December 2020).
- [16] Github, <https://github.com> (accessed in December 2020).
- [17] Google trends, <https://trends.google.com/trends> (accessed in December 2020).
- [18] Janez Brank, Gregor Leban, Marko Grobelnik. Annotating Documents with Relevant Wikipedia Concepts. Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017), Ljubljana, Slovenia, 9 October 2017.

Annex: DataBench Ontology Formalization

Namespaces

Schema	Prefix	Namespace
Schema.org	schema	http://schema.org
XML Schema	xsd	https://www.w3.org/2001/XMLSchema#
DBpedia Ontology	dbo	http://dbpedia.org/ontology/
DC Terms	dc	http://purl.org/dc/terms/
Databench terms	dbench	http://databench.ijs.si/ontology/
Fabio	fabio	http://purl.org/spar/fabio/
RDF Schema	rdfs	http://www.w3.org/2000/01/rdf-schema#

Classes

Class: Category	
Description	Category is used to group Topic and Tool under one class
SubClassOf	owl:Thing
Examples	<ul style="list-style-type: none"> - All Topics, e.g: Artificial Intelligence - All Tools, e.g. Apache_NLP

RDF	dbench:Category
Data Property: Title	
Description	The title of the entity
Domain	Category
Data Type	xsd:string
RDF	rdfs:label
Cardinality	1
Data Property: Wiki	
Description	Wiki link of category
Domain	Category
Data Type	xsd:anyURI
RDF	dbo:wikiPageWikiLink
Cardinality	1

Class: Topic	
Description	Topic is one type type of category that represents Topics in the area of Big data and Artificial intelligence
SubClassOf	Category
Examples	<ul style="list-style-type: none"> - Artificial Intelligence - Natural Language processing

RDF	dbench:Topic
Object Property: Sub Topic of	
Description	Hierarchical connection between topics and other topics
Domain	Topic
Range	Topic
RDF	dbench:SubTopicOf
Cardinality	0 or more

Class: Tool	
Description	Tool is one type type of category that represents the products, tools and libraries in the big data and artificial intelligence domain
SubClassOf	Category
Examples	<ul style="list-style-type: none"> - Apache NLP - postGreSQL
RDF	dbench:Tool
Object Property: Example of	
Description	Hierarchical connection between a tool and a topic
Domain	Tool
Range	Topic
RDF	dbench:ExampleOf

Cardinality	1 or more
--------------------	-----------

Class: Data Source	
Description	Instance consists of the the data source instances
Examples	<ul style="list-style-type: none"> - All research papers - All job postings - All news articles
SubClassOf	owl:Thing
RDF	dbench:DataSource
Data Property: Title	
Description	The title of the entity
Domain	Instance
Data Type	xsd:string
RDF	rdfs:label
Cardinality	1
Data Property: Creation Date	
Description	Date of publishing/creating
Domain	Data Source
Data Type	xsd:date

RDF	dc:created
Cardinality	1
Data Property: Description	
Description	The description of the entity, short summary, or abstract
Domain	Data Source
Data Type	xsd:string
RDF	dc:abstract
Cardinality	1
Data Property: URL	
Description	The url of the Data Source
Domain	Data Source
Data Type	xsd:anyURI
RDF	fabio:hasURL
Cardinality	1 to many
Data Property: Language	
Description	The language of the Data Source
Domain	Data Source
Data Type	xsd:language

RDF	dc:language
Cardinality	1
Data Property: Country	
Description	The countries involved in this Data Source
Domain	Data Source
Data Type	xsd:string
RDF	dbo:country
Cardinality	0 to many
Data Property: Source ID	
Description	The local Source ID
Domain	Data Source
Data Type	xsd:string
RDF	dc:identifier
Cardinality	1
Object Property: Entity of	
Description	Hierarchical connection between topics and other topics
Domain	Data Source
Range	Tool

RDF	dbench:EntityOf
Cardinality	1 or more

Class: Research Paper	
Description	Research is one type of Data Source that represents all Research Papers from Microsoft Academic Graph (MAG)
SubClassOf	Data Source
Data Source	http://academic.microsoft.com/
Examples	-
RDF	dbench:ResearchPaper
Data Property: Paper Type	
Description	The type of research paper
Domain	Research Paper
Data Type	xsd:string, restricted to "Journal", "Conference", "ArXiv", "Book", "Other"
RDF	dbench:paperType
Cardinality	1
Data Property: Affiliation	
Description	The affiliations of the authors of the research paper

Domain	Research Paper
Data Type	xsd:string
RDF	dbench:affiliation
Cardinality	0 to many
Data Property: Author	
Description	The authors of the research paper
Domain	Research Paper
Data Type	xsd:string
RDF	dc:creator
Cardinality	0 to many
Data Property: Publisher	
Description	The publisher of the research paper
Domain	Research Paper
Data Type	xsd:string
RDF	dc:publisher
Cardinality	0 or 1
Data Property: Journal	
Description	The name of the journal where the paper got published

Domain	Research Paper
Data Type	xsd:string
RDF	dbench:journal
Cardinality	0 or 1
Data Property: Conference	
Description	The name of the conference where the paper got published
Domain	Research Paper
Data Type	xsd:string
RDF	dbench:conference
Cardinality	0 or 1
Data Property: Citations	
Description	The number of citations of the research paper
Domain	Research Paper
Data Type	xsd:nonNegativeInteger
RDF	dbench:citations
Cardinality	1

Class: News Article

Description	News Article is one type of Data Source that represents all news article instances from Event Registry service
SubClassOf	Data Source
Data Source	http://eventregistry.org/
Examples	-
RDF	dbench:NewsArticle
Data Property: Time	
Description	The time of publishing
Domain	News Article
Data Type	xsd:time
RDF	dbench:createdTime
Cardinality	1
Data Property: News Source	
Description	The source of the news article
Domain	News Article
Data Type	xsd:string
RDF	dc:publisher
Cardinality	1
Data Property: sentiment	

Description	The sentiment of the news article
Domain	News Article
Data Type	xsd:decimal, restricted from 0 to 1
RDF	dbench:sentiment
Cardinality	0 or 1

Class: Job Posting	
Description	Job Posting is one type of Data Source that represent all job posting instances from Adzuna service
SubClassOf	Data Source
Data Source	https://www.adzuna.com/
Examples	-
RDF	dbench:JobPosting
Data Property: Salary Lower Limit	
Description	The lower limit in the salary interval of the job posting
Domain	Job Posting
Data Type	xsd:decimal
RDF	dbench:salaryLowerLimit

Cardinality	1
Data Property: Salary Upper Limit	
Description	The upper limit in the salary interval of the job posting
Domain	Job Posting
Data Type	xsd:decimal
RDF	dbench:salaryUpperLimit
Cardinality	1
Data Property: Salary Currency	
Description	The currency of the salary
Domain	Job Posting
Data Type	xsd:string, restricted to “dollars”, “euros”, “pounds”
RDF	dbench:currency
Cardinality	1
Data Property: Time	
Description	The time of publishing
Domain	News Article
Data Type	xsd:time
RDF	dbench:createdTime

Cardinality	1
--------------------	---

Class: Github Repository	
Description	Github repository is one type of Data Source that represent all github open repository instances
SubClassOf	Data Source
Data Source	http://github.com/
Examples	-
RDF	dbench:GithubRepo
Data Property: Forks	
Description	The number of forks from the repository
Domain	Github Repository
Data Type	xsd:nonNegativeInteger
RDF	dbench:forks
Cardinality	1
Data Property: Stars	
Description	The number of stars
Domain	Github Repository
Data Type	xsd:nonNegativeInteger

RDF	dbench:stars
Cardinality	1
Data Property: License	
Description	The type of license
Domain	Github Repository
Data Type	xsd:string
RDF	dc:license
Cardinality	0 or 1
Data Property: Programming Language	
Description	The programming languages used to implement the project
Domain	Github Repository
Data Type	xsd:string
RDF	dbench:programmingLanguage
Cardinality	0 to many
Data Property: Watchers	
Description	The number of watchers
Domain	Github Repository
Data Type	xsd:nonNegativeInteger

RDF	dbench:watchers
Cardinality	1
Data Property: PullRequests	
Description	The number of pull requests
Domain	Github Repository
Data Type	xsd:nonNegativeInteger
RDF	dbench:pullRequests
Cardinality	1
Data Property: Issues	
Description	The number of issues
Domain	Github Repository
Data Type	xsd:nonNegativeInteger
RDF	dbench:issues
Cardinality	1

Class: EU Research Project	
Description	EU research project is one
SubClassOf	Data Source
Data Source	https://cordis.europa.eu/
Examples	-
RDF	dbench:EUResearchProject
Data Property: Affiliation	
Description	The partner institutes in the eu research project
Domain	EU Research Project
Data Type	xsd:string
RDF	dbench:affiliation
Cardinality	0 to many
Data Property: Start Date	
Description	The starting date of the research project
Domain	EU Research Project
Data Type	xsd:date
RDF	dbench:startDate
Cardinality	0 or 1

Data Property: End Date	
Description	The ending date of the research project
Domain	EU Research Project
Data Type	xsd:date
RDF	dbench:endDate
Cardinality	0 or 1
Data Property: Coordinator	
Description	The coordinator of the project
Domain	EU Research Project
Data Type	xsd:string
RDF	dbench:coordinator
Cardinality	1
Data Property: Funding Program	
Description	The funding program of the project
Domain	EU Research Project
Data Type	xsd:string
RDF	dbench:fundingProgram
Cardinality	1

Data Property: Funding Schema	
Description	The mechanism in which the project is funded
Domain	EU Research Project
Data Type	xsd:string
RDF	dbench:fundingSchema
Cardinality	1
Data Property: Project Cost	
Description	The total cost of the project, in euros
Domain	EU Research Project
Data Type	xsd:nonNegativeInteger
RDF	dbench:cost
Cardinality	1
Data Property: Project Status	
Description	The status of the project
Domain	EU Research Project
Data Type	xsd:string, restricted to “Ongoing”, “Completed”, “Stopped”
RDF	dbench:status
Cardinality	1

Class: Stat	
Description	Statistic about a certain Category in a specific month
SubClassOf	owl:Thing
Examples	<ul style="list-style-type: none"> - Research Paper stat in 2016 - Job posting stat in 2020
RDF	dbench:stat
Object Property: Stat Category	
Description	The category the stat is about
Domain	Stat
Range	Category
RDF	dbench:statCategory
Cardinality	1
Data Property: Stat Data Source	
Description	The data source the stat is about
Domain	Stat
Data Type	xsd:string, restricted to “Research Paper”, “News Article”, “Github Repository”, “Job Posting”, “EU Research Project”
RDF	dbench:statDataSource
Cardinality	1

Data Property: year-month	
Description	The year and month of the stat
Domain	Stat
Data Type	xsd:gYearMonth
RDF	rdfs:yearMonth
Cardinality	1
Data Property: Stat Count	
Description	The count of the Data Source instances about the category in the specified month
Domain	Stat
Data Type	xsd:nonNegativeInteger
RDF	dbench:statCount
Cardinality	1