# There is no Data Science without Data Governance: a Proposal Based on Knowledge Graphs

Besim Bilalli[1], Petar Jovanovic[1], Sergi Nadal[1], Anna Queralt[1] and Oscar Romero[1,*]

[1]*Universitat Politècnica de Catalunya, UPC-BarcelonaTech*

### Abstract

Data Science and data-driven Artificial Intelligence are here to stay and they are expected to further transform the current global economy. From a technical point of view, there is an overall agreement that disciplines based on data require to combine data engineering and data analysis skills, but the fact is that data engineering is nowadays trailing and catching up with the rapid changes in the data analysis landscape. To unleash the real power of data, data-centric systems must be professionalized, i.e., operationalized and systematized, so that repetitive, time-consuming and error-prone tasks are automated. To such end, we propose our vision on next generation data governance for data-centric systems based on knowledge graphs. We claim that without the knowledge embedded in the data governance layer, Data Science will not unleash its potential.

### Keywords

data lifecycle, data management, data analytics, data governance, data science

## 1. A Data-Centric System

We are nowadays witnessing the raise of the so-called data-driven economy where data is an organization asset from where to extract objective evidences and gain competitiveness. However, all the promises related to data and its transforming aspects, are beyond realization.

First, collecting, organizing and managing large data repositories is hard. Concepts such as data lakes, data fabric, data mesh or DataOps, among many others, have arisen to help systematizing and operationalizing data management. Yet, current solutions require a huge manual burden and there are still no reference architectures (such as Data Warehousing for Business Intelligence, which is however not suitable for the problems framed by Data Science) [1]. Thus, organizations tend to work with different data silos, which are fragmented views of their own data that, in many cases, they are not able to cross. As a result, most data analysis conducted nowadays are based on certain available data, which are neither properly contextualized nor contain all the potentially relevant variables in the organization.

The main reason behind all these problems is the lack of governance of the whole data lifecycle. Data governance may be defined as *to what decisions must be made to ensure effective data management and data usage and who makes the decision* [2]. We identify the four main aspects required to govern the complete data lifecycle [3]: the *data principles* to establish the link between the data assets and the business, *machine readable metadata* to describe, not only the data assets, but also information about how to access and manipulate data. Metadata describing the complete data lifecycle within the system is mandatory (i.e., datasets used in a specific analysis, transformations and data preparation performed, algorithm chosen, model training information, etc.). Finally, a traversal but equally relevant aspect is *data quality*, which includes the qualitative description of the data assets. Importantly, as part of the metadata describing the data lifecycle, transformations conducted to guarantee data quality must be included.

In short, data governance claims for a systematic organization and annotation of data assets. Yet, current works either focus on how to organize data assets (i.e., data management) or to annotate it with metadata (data enrichment). But there are no end-to-end data governance proposals covering the whole data lifecycle.

Figure 1 presents the ambitious architectural framework we propose to make data governance true.

Our vision is grounded on four main subsystems: (i) the data management subsystem stores and manages the data assets, (ii) the data analysis subsystem is where the analytics take place, (iii) the data governance subsystem, where all the decisions, transformations and actions made at any step of the data lifecycle are annotated in a machine-readable format using knowledge graphs and (iv) the exploitation subsystem, where a set of modules, which interface the data governance subsystem, embed usual actions (e.g., create artifacts in the data management and / or analysis subsystems). As such, this architecture mimics that of a database system and, ideally, user interactions should always be conducted via the exploitation layer to guarantee that, whatever action taken, it

*DOLAP'24: 26th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data*

besim.bilalli@upc.edu (B. Bilalli); petar.jovanovic@upc.edu (P. Jovanovic); sergi.nadal@upc.edu (S. Nadal); anna.queralt@upc.edu (A. Queralt); oscar.romero@upc.edu (O. Romero)

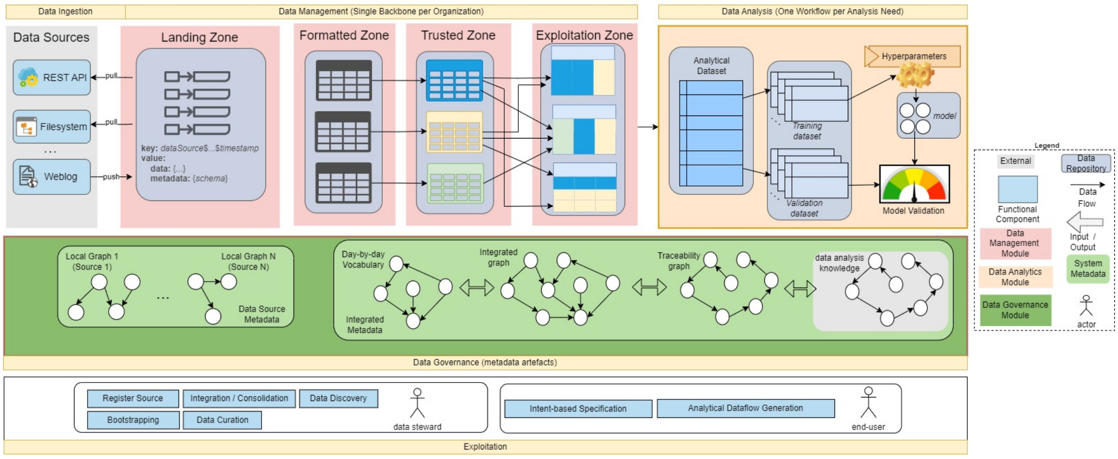

**Figure 1:** Our vision: Knowledge-graph governed data management and data analysis backbones

is properly annotated in the data governance subsystem (portraying the data independence principle).

Relevantly, the data management subsystem follows good practices and distributes the data assets (from raw data to other levels of transformations) into zones to separate concerns and facilitate maintenance and evolution. A dataset is registered into the system via the *register source* module. Registering a dataset automatically triggers several automatic tasks: (i) generate a graph-based representation of its schemata (also known as *bootstrapping*) and (ii) mappings (via the *data discovery* module) to a (iii) formatted representation of such data according to the chosen canonical data model (e.g., key-value). The *integration* module consolidates a set of datasets into a single *integrated graph*, which represents the system integrated schema. Relevantly, mappings between the integrated and local graphs allow to query the system via the integrated graph for exploration purposes. The integrated graph is the core metadata artifact through which the users will interact with the system. For example, data quality actions are conducted on top of the *integrated graph* (and propagated to the sources) via the *data curation* module, whose data assets are stored in the *trusted zone*. The *day-by-day vocabulary*, linked to the integrated graph, allows the users to express their needs in terms of their known vocabulary. Accordingly, end-users may express an analytical intent on top of the integrated graph via the *intent-based specification* module. This module leverages on the *analytical dataflow generation* module that first materializes an integrated dataset in the *exploitation zone* and then, from it, generates the required *data analysis workflow* according to the intents expressed. Finally, all decisions made during the execution of any of the modules mentioned is properly annotated in the *traceability graph*.

The core of this architecture is the layered knowledge-graph created for data governance, which will enable the development of next generation data-centric systems providing several benefits, specially, in the data analysis end, that will smooth current difficulties in data-centric projects. In short, we claim that a rigorous data governance: (i) facilitates **systematizing and operationalizing** data-centric projects, where data-related artifacts are organized to facilitate developing, maintaining and evolving complex operations on top of them; (ii) enables **automation** of complex processes. Specifically, we target the full automation of repetitive, time-consuming and error-prone tasks both for data management and analysis. Governance brings many benefits in this aspect: (a) the burden of collecting, storing and managing datasets is mostly hidden from the end-user, and (b) data analysis can be automated, in simple scenarios, via analytical intents expressed over the integrated graph. (c) Although we acknowledge that some aspects of the data lifecycle cannot be fully automated, these can be supported (e.g., rank alternatives): data integration, interpretation of analytical results, etc. Finally, governance (iii) generates **rich metadata** that can be analyzed to conduct meta-analysis about how data is used at any levels: collected, stored, transformed, analyzed, etc. or or use that knowledge to enrich / contextualize data analysis (e.g., to avoid LLMs hallucination).

## Acknowledgments

This work is supported by the Horizon Europe Programme under GA.101135513 (CYCLOPS) and GA. 101093164 (ExtremeXP) and the Spanish Ministerio de Ciencia e Innovación under project PID2020-117191RB-I00 / AEI/10.13039/501100011033 (DOGO4ML).

# References

[1] T. D. Bie, L. D. Raedt, J. Hernández-Orallo, H. H. Hoos, P. Smyth, C. K. I. Williams, Automating data science, Commun. ACM 65 (2022) 76–87.

[2] P. Weill, J. Ross, IT Governance: How Top Performers Manage IT Decision Rights for Superior Results, 2004.

[3] S. Nadal, P. Jovanovic, B. Bilalli, O. Romero, Operationalizing and automating data governance, J. Big Data 9 (2022) 117.