Open Data Product Specification

Jarkko Moilanen[1 [0000−0003−3470−2660]]

University of Jyväskylä, Finland

Abstract. Data is increasingly also becoming an article of trade or commerce and approached with a product mindset. The process and tools to create and publish data commodities in data marketplaces are based on scattered and provider-specific metadata models. Data consumers also find it hard to compare data products and reusable development of dataops software solutions is cumbersome. Hence, we propose an Open Data Product Specification which is a vendor-neutral, open-source machine-readable data product metadata model, which enables interoperability between organizations, data platforms, marketplaces, and tools. The specification is built on experiences gained from over 30 data product cases. Our artifact can be used by practitioners to increase the speed of designing, testing, implementation, and deployment of data products, and to speed up emerging data markets development.

Keywords: standard · data product · data economy.

1 Design of the artifact

Data-driven digitalization is proven to be profitable since according to a survey conducted by McKinsey Global Institute 2013 data-driven organizations are over 20 times more likely to acquire customers, half a dozen times as likely to retain their customers, and 19 times as likely to be profitable [1]. The digital transformation leads to major changes in established value creation structures and traditional business models of companies [7, 6]. Data are increasingly used beyond the improvement of internal processes by serving as a strategic resource for the development of data-driven innovations and business models [8, 9].

This data-driven innovation and creation of economic value is less and less created by a single organization or in traditional value chains but instead takes place in cross-industry, socio-technical networks – so-called data ecosystems [4, 5, 7]. Since data has to be used outside own organization, the reusability of it has to be increased to the next level. The era of data is about the process of data commoditization, where data is becoming an independently valuable asset that is freely available on the market. Given the nature of data ecosystems requiring border crossing activities in value creation, data used in the process must be packaged into products and services for more efficient reuse and sales. One holistic approach to increase the reusability and tradeability of data is to design and implement data products and services. It is suggested that this process of data commoditization is managed and led by data product managers. In short, data should be approached with product mindset [2].

The data products and data-driven service solutions are spread around an increasing amount of data marketplaces and described with a plethora of metadata models. This increases the amount of work done by data product managers since they need to handle multiple data product metadata models. One data source can be refined into multiple data products and services as described in the figure 1. The resulting data commodities are datasets, dashboards, API-driven data streams, and algorithms to mention a few. Often the targeted customers are in various marketplaces and to maximize the value derived from your data requires presence in multiple data marketplaces (see Figure 1). Without a common de facto standard on how to define data products, this work becomes very resource-consuming and error-prone.

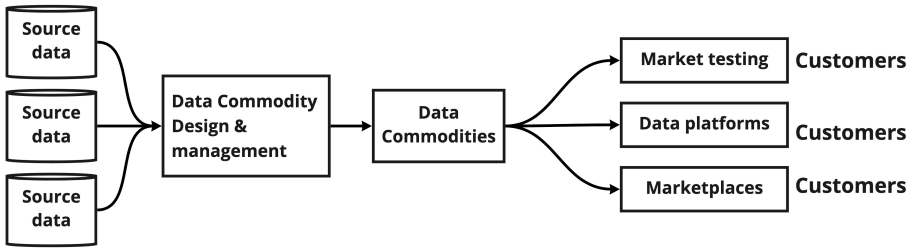

Fig. 1. Open Data Product Specification (ODPS) simplifies the process of creating multiple data products from one source and publishing data products in multiple data marketplaces, data platforms, and as mockups for market testing.

In addition, the tool stack for the data product design (basic tools for data product owners), development, and management is a wild west, consumers have a hard time knowing what they are purchasing or how to compare data products to find the best possible fit in their situation. This emerging data market mechanism is immature and scattered, which results in the cumbersome development of data-driven goods and services. In short, the data economy lacks a data product standard that would act as a baseline in the development, sharing, and comparison of data products.

The specification has four build-in aspects The Open Data Product Specification[1] is a vendor-neutral, open-source machine-readable data product metadata model. It defines the objects and attributes as well as the structure of digital data products. The work is based on existing standards (for example schema.org), best practices, and emerging concepts like Data Mesh[3], and over 30 data productizement cases in 13 companies over the past 2 years. In the name, the focus is on the latter words, and the prefix *open* refers to the openness of the standard. Any kind of connotations to open data are not intentional, intended,

[1] Open Data Product Specification version 1.0 https://opendataproducts.org/

or desirable. The specification aims to be a holistic approach to describe data products and other data commodities instead of limiting the approach to technical features only (see figure 2.). The word *product* is used in the name because in common business talk everything seems to be categorized as products even though some data commodities have more service features. This is expected to increase the familiarity and fit of the specification name is currently used rhetoric in business.

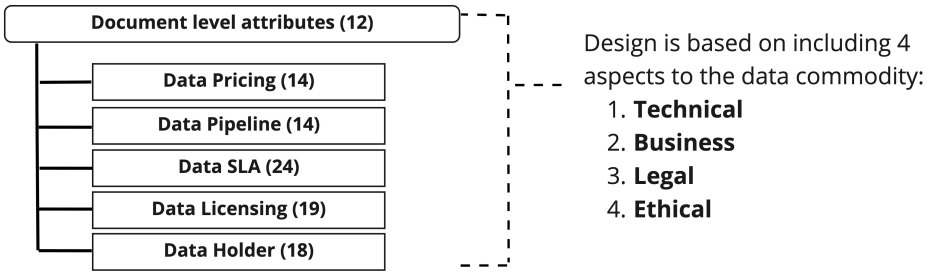

Fig. 2. Open Data Product Specification structure and in parenthesis amount of elements or attributes in each main objects.

The specification has been designed with four major aspects of the data product in mind: 1) technical (infrastructure & access), 2) business (pricing & plans), 3) legal (licensing & IPR), and 4) ethical (privacy & mydata). The four aspects are described in 5 elements, which contain attributes and other elements. Some of the attributes are mandatory while others are voluntary.

At the document level, the specification defines common product attributes such as name, product id, status, tags, value proposition, and visibility. In total document level has 5 mandatory attributes and 7 optional attributes. With help of the document level attributes data product designer can describe the data product basic information and status.

Next, we'll discuss the five main elements of the specification. The technical part of the data product is gathered inside *dataPipeline* element. Data Pipeline is a process whereby a data product pipeline deployment method is defined. Usually, the deployment script contains the logic of the individual steps as well as the code chaining the steps together. Data Pipeline object's purpose is enabling the building, deploying, and running the data product's code, and storing and giving access to data and metadata. This building principle has been adopted from the Data Mesh[3].

The business model of the data product is defined in *pricing* element. Pricing is the element whereby a business sets the price and conditions at which it will sell its products and services. The pricing object consists of mandatory and optional attributes. This element contains pricing plans related data to be used for example in displaying the items in a marketplace. If needed the standard metadata is converted to marketplace internal format. The pricing

plans are standardized and include the most common options: recurring time period based, one-time payments, pay-as-you-go, revenue sharing, data volume, dynamic pricing (high/low values), and pay what you want.

Data product SLA element contains attributes that define the desired and promised quality of the data product. This section of the specification describes the support functions, update frequency, service hours and methods, data product uptime and response times, monitoring services available for the consumer, and links to documentation and guides on how to utilize the commodity.

The Data Licensing element contains needed information to construct an agreement on the product usage. This section contains limitations of use including geographical limitation, reselling and modification rights, cancellation and continuity of the license, the URL of the Data Processing Agreement (DPA), and applicable laws. This part of the ODPS is the least standardized content-wise and consists of free text formatted fields. In the long run, the aim is to go towards the model used in Creative Commons.

The Data Holder section contains fields to describe the entity legally allowed to create, develop and publish data products. Data Holder concept has been adopted from Data Governance Act[2] in which it is defined as a legal person, public body, international organization, or a natural person who is not a data subject concerning the specific data in question, which, under applicable Union or national law, has the right to grant access to or to share certain personal data or non-personal data.

While the Open Data Product Specification tries to accommodate most use cases, additional data can be added to extend the specification at certain points. The extension properties are implemented as patterned fields that are always prefixed by "x-". The extensions may or may not be supported by the available tooling, but those may be extended as well to add requested support.

2 Significance to research

So far the discussions of data products and other data commodities have focused on rather technical aspects of data-driven value creation. The innovations have happened in technical solutions such as big data management, data lakes, and warehouses. Much of the innovations are focused on data reuse instead of monetization via marketplaces which requires crossing the company border. As far as the author knows, this kind of more holistic attempt to describe data products has not been done before in machine-readable format.

The model creates the first concise model for describing data products in a machine-readable format so that it includes business model information, legal conditions, quality, and SLA details as well as technical data pipeline details. It is the first attempt to standardize data commodity, metadata model. ODPS creates a shared understanding of what elements and concepts data product consists of and offers one approach to it.

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767

3 Significance to practice

The ODPS aims for the same impact in the Data Economy as what OpenAPI specification[3] did for the API Economy. In the API economy, OpenAPI specification standardized how REST APIs are described in a machine-readable format. It enabled faster and distributed tools development in API design, testing, and implementation, increased the discoverability of APIs, fast and automated mocking of API products for customer testing, and API documentation software development.

The ODPS has multiple benefits. Firstly, at a high level, the specification enables interoperability between organizations, data platforms, marketplaces, and tools. It creates the foundation for joint development and service pipelines as well as management solutions at the ecosystem level. ODPS is intended to be a de facto standard like OpenAPI. International standards are a vital tool in ensuring products and services are interchangeable and compatible across borders, removing trade barriers, reducing production and supply chain costs and building confidence in business services, and protecting consumers.

Secondly, it reduces data product metadata conversions and errors between systems and organizations. Currently, data marketplaces have their metadata models, which often overlap with each other only for a few attributes. In some cases, the metadata model is not even publicly available. This in turn makes each marketplace unique and the differences of metadata models between various participants in the data value chain lead to multiple conversions between the formats. In addition, the conversion rules and solutions must be developed, tested, and maintained. All that causes more costs and increases the risk of errors in the process. The mentioned limitations for data markets to blossom would be reduced significantly given that ODPS would be widely accepted and used.

Thirdly, ODPS can increase the speed of designing, testing, implementation, and deployment of data products. If the data product design is machine-readable then it can be used in an automatic generation on software tests, mock the data products in marketplaces, and create data pipeline automation (eg DataOps) as part of data product deployment.

Fourthly, ODPS can speed up tools development around data product design, development, and management. In the API Economy, the tool stack for API design, testing, and documentation exploded after the introduction and adoption of Swagger. Since ODPS is shared with an open license (Creative Commons BY-SA 4.0) it encourages the adoption and free use of it as part of tools and interoperability development.

4 Evaluation of the artifact

The specification has been developed based on experiences gained from 31 data product cases during 2020-2021. The author has been involved in most of those

[3] https://www.openapis.org/

and documented the findings on the fly. At the end of 2021, it was decided that enough cases has been done since no more new issues emerged. Version 1.0 of the Open Data Product Specification was written during Dec 2021 - Jan 2022. Before publishing the production version, a release candidate was pushed to 8 data economy professionals for review. The feedback was inspiring and positive. One of the reviewers was a technical lead of data platforms and according to him: "*Standardization of commonly needed parts. There is a clear need for this kind of standardization.*" A principal software developer emphasized the comprehensive nature of the specification: "*great way to keep everything together and inform about the data product and licensing model with it.*"

To get a better understanding of the specification in practical terms, a complete hello world type of example with the majority of the attributes used is available on the specification homepage.[4] Two organizations are known to do the first implementations of data pipeline solutions which will utilize the ODPS, but results from those cases are not yet available.

References

1. Bokman, A., Fiedler, L., Perrey, J., Pickersgill, A.: Five facts: How customer analytics boosts corporate performance | McKinsey (2021), https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/five-facts-how-customer-analytics-boosts-corporate-performance
2. Davis, J., Nussbaum, D., Troyanos, K.: Approach Your Data with a Product Mindset. Harvard Business Review (2020), https://hbr.org/2020/05/approach-your-data-with-a-product-mindset
3. Dehghani, Z.: Data Mesh: Delivering Data-Driven Value at Scale. O'Reilly Media, Inc., Farnham (2022)
4. Hein, A., Weking, J., Schreieck, M., Wiesche, M., Böhm, M., Krcmar, H.: Value co-creation practices in business-to-business platform ecosystems. Electronic Markets **29**(3), 503–518 (2019), publisher: Springer
5. Oliveira, M.I.S., Lóscio, B.F.: What is a data ecosystem? In: Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age. pp. 1–9 (2018)
6. Weill, P., Woerner, S.L.: Thriving in an increasingly digital ecosystem. MIT Sloan Management Review **56**(4), 27 (2015), publisher: Massachusetts Institute of Technology, Cambridge, MA
7. Yoo, Y., Henfridsson, O., Lyytinen, K.: Research commentary—the new organizing logic of digital innovation: an agenda for information systems research. Information systems research **21**(4), 724–735 (2010), publisher: INFORMS
8. Zakari, I.S.: Promoting Statistics in the Era of Data Science and Data-Driven Innovations. Statistics Education Research Journal **19**(1) (2020)
9. Zolnowski, A., Christiansen, T., Gudat, J.: Business model transformation patterns of data-driven innovations (2016)

[4] https://opendataproducts.org/#hello-world-example