Towards Modular Data Marketplaces

Soulmaz Gheisari[1], Semih Yumusak[1], Jaime Osvaldo Salas[1], Luis-Daniel Ibáñez[1], George Konstantinidis[1] and Dumitru Roman[2]

[1]*Department of Electronics and Computer Science, University of Southampton, Southampton, UK*
[2]*SINTEF, Oslo, Norway*

Abstract

Building a monolithic data marketplace is challenging due to complex inter dependencies, leading to cumbersome and error-prone development where a single failure can disrupt the entire system. To address this, we propose a modular approach using dynamic plugins in the UPCAST project. Our flexible framework allows components to be activated or deactivated as needed, enhancing scalability and resilience. By decoupling functionalities into interchangeable modules, we mitigate the risk of single points of failure, simplify maintenance, and facilitate customization for more robust marketplace solutions.

Keywords

Data consumer, Data marketplace, Data provider, Negotiation, Privacy and usage control, Resource specification, Resource discovery

1. Introduction

The UPCAST project[1] offers a set of plugins designed to automate data sharing and processing in data marketplaces, facilitating interactions between data consumers and data providers. We outline several key components relevant to UPCAST's plugins (see workflow in Figure 1.). The process begins with the Data Provider defining the resource specification, detailing the attributes, capabilities, and constraints of the resource. Next, the Data Consumer examines these specifications through a discovery process to identify potential resources that meet their needs[1]. Upon finding a suitable resource, the consumer generates a request to access it. This request undergoes a review process to ensure all privacy and access control criteria are satisfied. In cases where conflicts are detected, they must be resolved. The request is then sent to the provider, initiating a negotiation process. Once an agreement is reached, an UPCAST contract is generated and signed by both the provider and the consumer, finalizing the agreement.

1.1. Resource Specification Plugin

Data sources are annotated as a dcat:Dataset, with the data model designed as a knowledge graph using both the DCAT[2] and UPCAST vocabularies. Users initiate this plugin to specify the details required to create a new resource. The creation of a new resource involves the following sub-procedures:

- Import UPCAST vocabulary and domain-specific vocabulary in machine-readable format;
- Define metadata of the resource;
- Define access and usage policies of the resource;
- Assign energy profile to the resource that will be used to optimise the environmental impact;
- Associate price to the resource for further negotiations;
- Create resource profile/summary.

RuleML+RR'24: Companion Proceedings of the 8th International Joint Conference on Rules and Reasoning, September 16–22, 2024, Bucharest, Romania

s.gheisari@soton.ac.uk (S. Gheisari); semih.yumusak@soton.ac.uk (S. Yumusak); j.o.salas@soton.ac.uk (J. O. Salas); l.d.ibanez@soton.ac.uk (L. Ibáñez); g.konstantinidis@soton.ac.uk (G. Konstantinidis); dumitru.roman@sintef.no (D. Roman)
0000-0001-8974-2841 (S. Gheisari); 0000-0002-8878-4991 (S. Yumusak); 0000-0002-9353-8955 (J. O. Salas); 0000-0001-6993-0001 (L. Ibáñez); 0000-0002-3962-9303 (G. Konstantinidis); 0000-0001-6397-3705 (D. Roman)

[1] https://www.upcast-project.eu
[2] https://www.w3.org/TR/vocab-dcat-3/

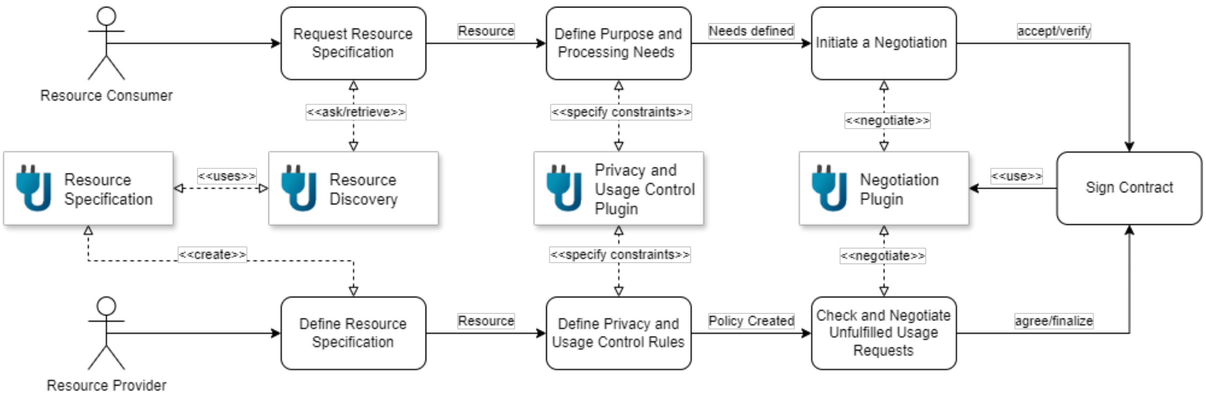

Figure 1: The Workflow in a Modular Data Marketplace.

1.1.1. Semantic Profiling

The data profiling service generates a profile for a dataset using a specified profiler given dataset metadata, sample data or the whole dataset, and other supplementary materials. A number of profilers can be connected to provide the "plug-and-play" profiling service according to the needs and requirements of the user, for example, profilers that give statistics on the dataset or provide semantic information about the data. In UPCAST, the main purpose of the profiling service is to enhance the representation of data to improve data discoverability, in particular, through semantic profiling.

1.2. Resource Discovery Plugin

The Resource Discovery Plugin acts as an intermediary, facilitating the retrieval of resource specifications. Resource consumers can request and retrieve information from the available resources provided by various providers. While searching the knowledge base, users may also find similar sources through semantic similarity search[2]. Therefore, resource discovery provides the following functionalities for a consumer:

- A comprehensive search for resources based on the consumer's intentions;
- Browsing for resources, offering the user an intuitive and efficient way to navigate and explore the available resources;
- Discovering related/recommended resources, ensuring up-to-date and dynamic results. The relevant resources graph is continuously updated as new datasets arrive.

Figure 2 illustrates the data model for both resource specification and resource discovery. This model details the structure and attributes necessary for specifying resources and discovering them within the UPCAST.

1.3. Privacy and Usage Control Plugin

After the resource specification, the resource provider defines constraints on the resources using Open Digital Rights Language (ODRL)[3] rules, leveraging both the UPCAST and domain-specific vocabularies. On the other hand, the resource consumer specifies the intentions via a Data Processing Workflow (DPW) specification and outlines any organisation-specific access and usage control rules, as well as rules prescribed by applicable regulations (e.g., GDPR). Subsequently, conflict identification occurs between the provider's constraints, the consumer's intentions, and internal rules, making the derivation of authorisation decisions possible. The functionalities of the plugin can be summarised as below:

- Transform the resource provider constraints to privacy and usage control rules;

[3] https://www.w3.org/TR/2018/REC-odrl-model-20180215/

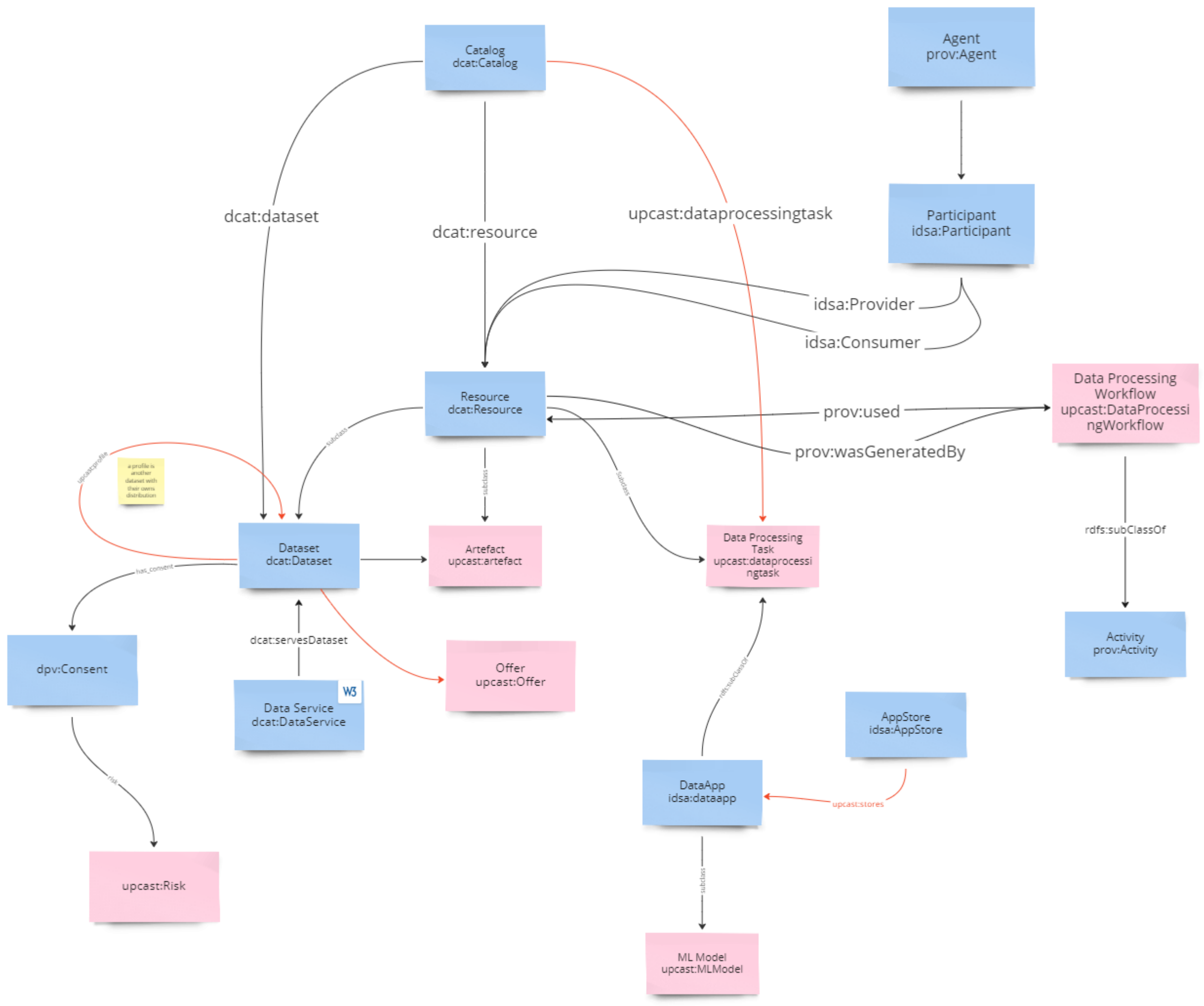

Figure 2: Resource Specification and Discovery Data Model

- Define rules for the resource consumer;
- Manage rules;
- Identify conflicts between the provider's constraints and the consumer's intentions;
- Access and usage decision making.

Figure 3 shows the data model of this plugin.

1.4. Negotiation Plugin

Often, the processing intentions of a data consumer for a dataset of their interest differ from what the data provider is willing to allow. These differences may include the purpose of the processing, the time interval for which the provider is willing to allow access, or the price to pay. Nevertheless, these differences are not necessarily irreconcilable, and both parties can often reach an agreement through negotiation [3]. The Negotiation and Contracting plugin within UPCAST, serves as a pivotal component, streamlining the complex processes of negotiation and contract management. With its multifaceted functionality, this plugin facilitates efficient communication and collaboration between data producers and consumers. First, the plugin provides a Policy Administration Point with a user-friendly graphical interface, enabling users to define restrictions, privacy, and usage policies in a user-friendly and intuitive manner. In addition, the Negotiation Plugin serves as a Policy Management Point (PMP) for usage

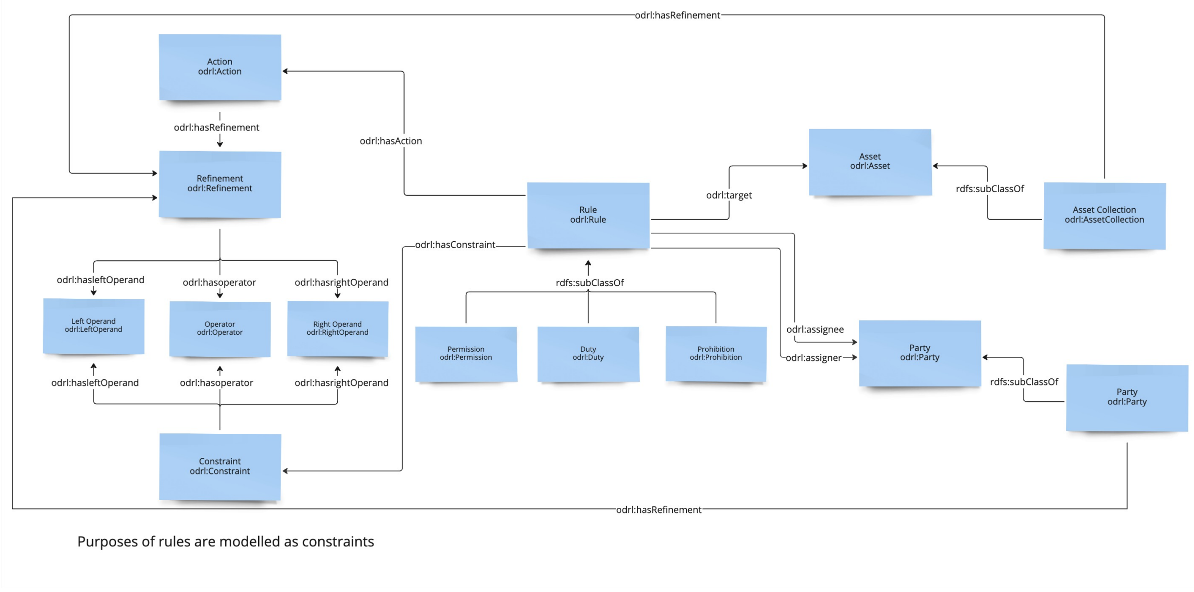

Figure 3: Privacy and Usage Control Data Model

restrictions by reading machine-readable policies and checking them against information from the privacy and usage control, environmental impact, and pricing plugins, and automatically reaching an agreement if there are no policy conflicts. Otherwise, if conflicts are detected, a negotiation will be initiated, allowing the data provider or consumer to present counteroffers. Figure 4 illustrates the negotiation and contracting plugin flowchart.

Upon the initiation of a negotiation process, the plugin provides a centralised platform for discussing terms, pricing, and specifications, allowing users to track, and finalise negotiations seamlessly. Moreover, the plugin incorporates robust contract management features, allowing users to create, review, and execute contracts with ease. By automating routine tasks and offering customisable Data Processing Workflows (DPWs), it enhances data sharing while ensuring compliance with regulatory requirements.

The provider will ultimately decide the negotiation's outcome by agreeing, rejecting, or sending another counteroffer. The result of a successful negotiation process is a data sharing contract [4] that extends the usage control specification defined by the International-Data-Spaces-Association (IDSA) [4], which in turn uses ODRL. Contracts also utilise other ontologies such as the Data Privacy Vocabulary (DPV)[5], which defines an ontology that allows for the definition of the use, processing and purpose of processing of data under relevant legislation, notably the GDPR, enabling more descriptive and technology-independent contracts.

1.4.1. Contract Generation Supported by LLM

The contract generation process within the UPCAST plugin is significantly enhanced by the integration of Large Language Models (LLMs). These advanced AI models facilitate the automatic generation of comprehensive and precise contracts based on the negotiation outcomes. By analyzing the details of the negotiation, including usage policies, pricing structures, and specific data processing requirements, the LLM can draft contracts that accurately reflect the agreed terms. This automation not only speeds up the contract creation process but also reduces the risk of human error and ensures that all legal and regulatory aspects are meticulously addressed. The LLM's ability to understand and generate natural language makes it an invaluable tool for creating clear and enforceable contracts, thereby streamlining the entire negotiation and contracting workflow within the UPCAST platform.

[4] https://internationaldataspaces.org/
[5] https://w3c.github.io/dpv/dpv/

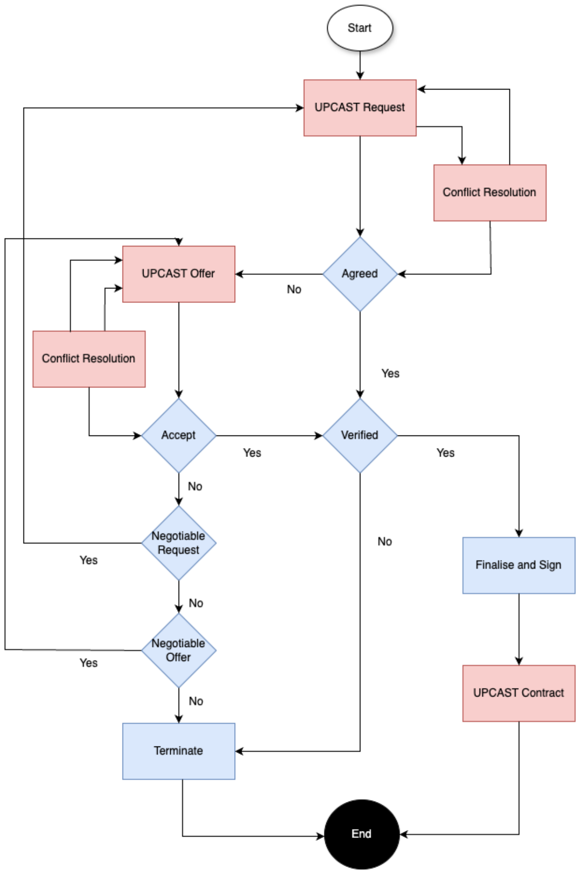

Figure 4: Negotiation and Contracting Flow Chart

2. Conclusion

In conclusion, this paper has introduced a modular approach to building data marketplaces, addressing the challenges posed by traditional monolithic systems. By utilising dynamic plugins within the UPCAST project, our solution provides a flexible framework that enhances scalability, resilience, and ease of maintenance. The decoupling of functionalities into discrete modules mitigates the risk of single points of failure and allows for tailored customisation to meet specific marketplace needs. This approach not only simplifies system upgrades and maintenance but also ensures robust and adaptable data marketplace solutions, demonstrating significant advantages over conventional monolithic designs.

Acknowledgments

This work was funded by the UKRI Horizon Europe guarantee funding scheme for the Horizon Europe projects UPCAST (101093216101093216) and RAISE (101093216101058479).

References

[1] G. Konstantinidis, L.-D. Ibáñez, D. Roman, Data marketplaces in the ai economy, in: Symposium on AI, Data and Digitalization (SAIDD 2023), 2023, p. 38.

[2] R. Sharifpour, M. Wu, X. Zhang, Large-scale analysis of query logs to profile users for dataset search, Journal of Documentation 79 (2023) 66–85.

[3] W. Fox, Y. Dautaj, International commercial agreements, Kluwer Law International BV, 2023.

[4] J. J. Chen, Multicenter observational studies: Understanding the basics of data sharing and data user agreements, 2024.