



Universidad Politécnica
de Madrid

**Escuela Técnica Superior de
Ingenieros Informáticos**



Master in Data Science

Master Thesis

Regulation Assessment for Data Spaces

Author: ChiaChi Hsu

Madrid, July, 2024

This Master Thesis has been deposited in ETSI Informáticos de la Universidad Politécnica de Madrid.

Master Thesis

Master in Data Science

Title: Regulation Assessment for Data Spaces
July, 2024

Author: ChiaChi Hsu

Supervisor: Víctor Rodríguez Doncel, Carlos Ruiz Moreno
Department of Artificial Intelligence
ETSI Informáticos
Universidad Politécnica de Madrid

Abstract

As data spaces become increasingly important in the European data market, more companies and individuals need to focus on complying with the regulations established by the European Union. However, since these regulations have only been developed in recent years and are complex, it is difficult for most companies and individuals to determine whether their data-related processes comply with the relevant regulations. Therefore, the goal of this project is to develop a tool presented as a wizard-like website to help users assess whether their data complies with regulations. A proof-of-concept tool has been implemented to simplify the compliance assessment process and reduce the risk of noncompliance by providing specific functions and guidance on ensuring data compliance. Several legal scenarios have been modeled to test and evaluate the tool's effectiveness. Through this website, users, especially non-legal experts, can easily fulfill regulation compliance tasks.

Resumen

A medida que los espacios de datos se vuelven cada vez más importantes en el mercado de datos europeo, más empresas e individuos necesitan centrarse en cumplir con las regulaciones establecidas por la Unión Europea. Sin embargo, dado que estas regulaciones solo se han desarrollado en los últimos años y son complejas, es difícil para la mayoría de las empresas e individuos determinar si sus procesos relacionados con datos cumplen con las regulaciones pertinentes. Por lo tanto, el objetivo de este proyecto es desarrollar una herramienta presentada como un sitio web similar a un asistente para ayudar a los usuarios a evaluar si sus datos cumplen con las regulaciones. Se ha implementado una herramienta de prueba de concepto para simplificar el proceso de evaluación de cumplimiento y reducir el riesgo de incumplimiento al proporcionar funciones específicas y orientación sobre cómo asegurar el cumplimiento de los datos. Se han modelado varios escenarios legales para probar y evaluar la efectividad de la herramienta. A través de este sitio web, los usuarios, especialmente los no expertos en leyes, pueden cumplir fácilmente con las tareas de cumplimiento de regulaciones.

Contents

1 Introduction	1
1.1 Motivation	1
1.2 Context	1
1.2.1 Data Spaces Participants	2
1.2.2 European Standards	3
1.2.2.1 Data intermediation services	3
1.2.2.2 Data Altruism	4
1.2.2.3 Data Act	5
1.3 Objective	5
2 State of the Art	6
2.1 BPMN and BPEL	6
2.1.1 Encoding norms in BPMN	7
2.1.1.1 Use case 1: eGovernment service	7
2.1.1.2 Use case 2: Data Breach	9
2.1.2 BPMN 2.0 diagram	11
2.1.2.1 Elements and Symbols	11
2.2 Tools for Regulation Assessment	12
2.2.1 PET Explorer	12
2.2.2 High-Risk AI	12
2.2.3 Data Protection Self Assessment	14
2.2.4 ValidaCripto RGPD	14
3 Analysis	17
3.1 Scenarios	17
3.1.1 Copyright Scenario	17
3.1.1.1 Copyright Scenario - Method II	22
3.1.1.2 Data Spaces Scenario	24
3.2 Requirements	27
4 Results and conclusions	28
4.1 Implementation	28
4.1.1 Architecture	28
4.1.1.1 Frontend	28
4.1.1.2 Backend	30
4.2 Evaluation	31
4.3 Contributions	32
4.4 Conclusions	32

4.5 Future Work	32
Bibliography	34
Annex	35

List of Figures

2.1	BPMN modelling of an eGovernment service	8
2.2	BPMN model for pattern Data Breach	10
2.3	High-risk conditions using the 5 concepts	13
2.4	User interface of ValidaCripto RGPD	15
2.5	Elements under the component KEYS	15
2.6	Validation section	16
3.1	BPMN diagram of "Copyright" scenario	20
3.2	Demo of the question in "Copyright" scenario	22
3.3	BPMN diagram of "Copyright_text" scenario	23
3.4	Demo of the question in "Copyright_text" scenario	24
3.5	BPMN diagram of "Data Spaces" scenario	26
3.6	Demo of the question in "Data Spaces" scenario	27
4.1	High-level system diagram	28
1	Introduction of the project	35
2	Example of the application	35
3	App Instruction (Local)	36
4	App Instruction (Online)	37
5	Acknowledgements and Disclaimer	37

List of Tables

2.1 Symbols and their usage 11

Listings

3.1 BPEL Excerpt 21

Chapter 1

Introduction

1.1 Motivation

The aim of the project is to develop a tool to simplified data compliance tasks. In the framework of Common European Data Spaces, data needs to be assessed in terms of its compliance with new relevant EU regulations and policies. The tool would provide comprehensive reports and recommendations to help organizations ensure that their data spaces meet the necessary regulatory standards.

This project is driven by the idea of Rules as Code (RaC). Rules as Code is a concept that involves converting government rules, such as legislation, regulations, standards, and policies, into code to make them machine-readable and consumable. Coded rules could bring clarity and transparency to a rule by converting complicated legal and legislative language into simple to understand logical sequences. This could make rules simpler to understand and interpret, improving the quality of the rule. Additionally, with rules written as code, government creates transparency, trust and open engagement with citizens.[1]

1.2 Context

GAIA-X is an European initiative that aims to establish a federated data infrastructure and a data sharing system to promote data sovereignty, interoperability, and trust in the digital ecosystem. Data spaces is a fundamental part of GAIA-X, indicating a virtual space focus on creating a secure and transparent environment for data sharing and collaboration among participants.

The idea of data spaces was introduced as an approach to data integration where focus on integrating data at a semantic level using shared vocabularies, without the need for a common database schema.

The core concept of data spaces revolves around creating a distributed data integration framework that enables data sharing, collaboration, and innovation within ecosystems. Data space functions as a market of data, where different contributors act like various stores in this market, freely sharing and exchanging data according to the European standard. This facilitates the easy use and analysis of data, thereby

creating value from it.

Key features of the Common European Data Spaces according to European Commission [2]:

- Open for all participants (organizations and individuals).
- Infrastructure is secured to pool, access, share, process, and use data.
- Respect EU rules, especially personal, consumer protection, and competition law.
- Enable data holders to grant access to data or to share certain personal or non-personal data.
- Allow data holders to make their data available for freely reuse.

Data Spaces can apply in various domains such as healthcare, finance, industrial, agricultural, energy, telecommunications, transportation, and smart cities. Data Spaces enable organizations to leverage data effectively for decision-making, innovation, and value creation processes. For instance, in the medical sector, data spaces can secure the sharing of patient data for personalized medical treatment. In manufacturing, data spaces optimize supply chain operations and enhance productivity through real-time data sharing. In transportation, data spaces enable route optimization, traffic management, and passenger services for enhanced efficiency and safety. In conclusion, the growth of data spaces is proportional to the number of participants, so it is beneficial to encourage organizations to collaborate, innovate, and discover insights from shared data. [3]

1.2.1 Data Spaces Participants

According to datos.gob.es¹, there are four types of participants in a data space:

- **Data space promoters:**
In charge of sharing, operating, and managing data spaces, as well as attracting new participants. This type of participant is fundamental in expanding the potential of data spaces. As more participants join in to develop data spaces, more business models can be generated based on the concept of data spaces.
- **Data and service providers:**
Provide data sets and service. data providers have various options of revenue generate model. For example, data providers can offer free data, using it to attract users and thereby draw in advertisers to generate revenue. Alternatively, they can charge for specific services or data beyond the free data.
- **Consumers:**
Purchase data products from data spaces. They can use this third-party data to generate information and solve business problems that would be challenging to address without additional data. Additionally, they can process the obtained information and offer it back into the data space as value-added services.
- **Technology roviders:**
In charge of deploying and operating essential technologies that shape the data-

¹<https://datos.gob.es/en/blog/why-data-spaces>

Introduction

sharing ecosystem. Their work includes development, configuration, parameterization, deployment, governance, and management of the data space.

1.2.2 European Standards

As the amount of data is significantly increasing, data researchers can use it to create more economic and social value, including but not limited to benefits in fields such as healthcare, engineering, and innovation. However, due to technological barriers and issues of trust in data sharing, this valuable data has not yet been maximized in application. Therefore, to eliminate these problems and promote data sharing and reuse, the DGA (Data Governance Act) regulations entered into force on 23 June 2022 and, is applicable since September 2023.

The DGA aims to regulate the re-use of protected data, by boosting data sharing through data intermediate services and by promoting the idea of data altruism. The DGA protects both personal and non-personal data while GDPR applies wherever personal data is concerned.

1.2.2.1 Data intermediation services

One of the key goals of DGA that related to Data Spaces is to establish data intermediation services. Many companies worry that sharing their data could lead to economic losses or a decrease in market competitiveness, leading to reservations about data sharing. Therefore, the DGA has established a set of rules to enhance the trust of these data providers in data sharing. According to European Commission [4], data intermediaries will function as neutral third parties that connect individuals and companies with data users. While they may charge for facilitating the data sharing between the parties, they cannot directly use the data that they intermediate for financial profit. Data intermediaries need to remain neutral in this process, avoiding conflicts of interest, and must be structurally separate from other services or institutions to ensure their neutrality.

A data intermediate service that already put into practice is **Dawex**² which define itself as the *leader in data exchange solutions to distribute or share data products, with trust, for any business case, in compliance with data regulations*. Instead of purchasing vast amounts of data from organizations and selling it themselves, Dawex acts as a bridge that links data providers and data acquirers, allowing both participants to conduct their data sharing and acquiring activities. Here are examples of data ecosystems and data spaces created with Dawex's technology:

- **Corporate Data Hub**³:

This service helps multinational enterprises break data silos across their organizations by enabling data providers and data acquirers to publish and search for data products internally. It is particularly beneficial to organizations by providing a better understanding of the market and further improving operations.

- **Data Marketplace**⁴:

A Data Marketplace, or Data Shop when there is only one data provider, provides

²<https://www.dawex.com/en/>

³<https://www.dawex.com/en/solutions/corporate-data-hub/>

⁴<https://www.dawex.com/en/solutions/data-marketplace-data-shop/>

a platform for data providers to sell valuable data products to acquirers, generating revenue. Providers control usage conditions and access rights, ensuring no loss of benefits from sharing their data. This fosters data-sharing relationships within organizations and partners, enhancing market position. The marketplace can also expand into a larger Industry Data Hub.

- **Industry Data Hub** ⁵:

The Industry Data Hub allows several partner organizations in the same industry to share data products to achieve common business or industrial objectives. Data providers can choose whether to share data products for free or commercialize them. Various benefits come with the Industry Data Hub, such as optimization of operations and customer experience, and acceleration of building information modeling for smart cities. **Agdatahub** ⁶ is an example of an industry data hub that focuses on intermediating agricultural data.

1.2.2.2 Data Altruism

According to Article 2(16) Data Governance Act (DGA) ⁷, " *'data altruism' means the voluntary sharing of data on the basis of the consent of data subjects to process personal data pertaining to them, or permissions of data holders to allow the use of their non-personal data without seeking or receiving a reward that goes beyond compensation related to the costs that they incur where they make their data available for objectives of general interest as provided for in national law, where applicable, such as healthcare, combating climate change, improving mobility, facilitating the development, production and dissemination of official statistics, improving the provision of public services, public policy making or scientific research purposes in the general interest.*"

In practical application, organizations that endorse data altruism and provide data can be registered as "data altruism organisations recognised in the Union". They must operate on a non-profit basis, be transparent, and have measures in place to safeguard the data sharing rights of citizens and companies. Additionally, they need to comply with relevant regulations concerning information, technology, security, and other aspects. The European Union's common Data Altruism Agreement is structured in a standardized format to collect data from member states. This enables entities sharing data to easily give and withdraw their consent.

Due to health data being spread across different countries and regulated differently, the Towards European Health Data Space (TEHDAS) ⁸ project aims to develop European principles for the secondary use of health data. The secondary use ranges from research, decision-making, development, innovation, and education. This facilitates personalized medical treatment and boosts healthcare services, ultimately improving overall human health. To achieve this goal, TEHDAS encourages individuals and organizations to share their health data based on data altruism, allowing researchers access to vast amounts of data for research and innovation purposes.

⁵<https://www.dawex.com/en/solutions/industry-data-hub/>

⁶<https://agdatahub.eu/en/entreprise/>

⁷https://www.european-data-governance-act.com/Data_Governance_Act_Article_2.html

⁸<https://tehdas.eu/>

1.2.2.3 Data Act

Data Act is another European regulation related to Data Space, emphasising fair access and user rights, while ensuring the protection of personal data. ⁹ The Data Act, as a complementary legislation to the Data Governance Act (DGA), focuses more on data sharing and circulation. Its main objectives include enhancing the sharing of data between different organizations to promote innovation and economic growth. The Data Act also ensures the fairness and transparency of data usage, preventing unfair competition and benefiting small and medium-sized enterprises and startups in their data utilization. Additionally, the Data Act ensures the free flow of data within the EU and internationally, preventing localization restrictions, which is a crucial characteristic of the Data Space.

1.3 Objective

This project aims to develop a method that simplifies the process of data compliance by facilitating the easy creation of workflow process diagrams that reflect specific legal aspects, and their automated transformation into wizard-like web applications particularly suitable for non-legal experts. Specifically, the objectives are:

- O1. To encode at least two real legal compliance processes in a formal manner, using BPMN as a well-known, standard means of formalizing the process.
- O2. To demonstrate the approach by programming a demo that transforms the processes above into an actionable tool that provides legal insights to laymen. The demonstrator will achieve a Technology Readiness Level (TRL) of 3.

⁹<https://digital-strategy.ec.europa.eu/en/policies/data-act>

Chapter 2

State of the Art

In this chapter, we will justify the technological choices and the related elements that will be used during development. In section 2.1, we will introduce BPMN and BPEL as our technological choices along with the elements that will be used in this work. Next, in section 2.2, four alternative approaches to tackle similar problems are also mentioned to suggest potential methods for achieving our goal.

2.1 BPMN and BPEL

Business Process Modeling Notation (BPMN) helps improve efficiency in business by visualizing business processes in various fields. As an international standard, it is well-known among professionals and mature enough to support a wide range of needs. The application of this project will be presented in the form of a process, classifying the appropriate judgment results based on different scenarios. Therefore, the concept of BPMN is suitable to develop this software. The BPMN diagram makes the business process more understandable to all stakeholders. BPMN diagrams allow different stakeholders to visualize business processes, making it easier to make workflows more effective and efficient[5]. For management-level participants, it enables them to grasp the planned business process at a glance. At the implementation level, BPMN provides the standard process to which they should adhere. Thus, BPMN is useful for both technical and non-technical sides, potentially avoiding any misunderstandings or misleading messages during the implementation of the business process.

Business Process Execution Language (BPEL) is an XML-based language that can automate the implementation of processes developed in BPMN for technical analysts and programmers. BPEL is an interoperable, portable language for defining business processes, with potential applications in quality of service, human activities, grid computing, and autonomic computing[6]. BPEL is parseable and easy to execute with tools like Camunda Modeler ¹ and bpmn.io ². Using Camunda and bpmn.io is a standard way to produce BPMN diagrams. The output from both tools can be handled in the same way, making them interchangeable solutions. Both of them provide

¹<https://camunda.com/bpmn/>

²<https://bpmn.io/>

an environment for editing BPMN diagrams and managing BPEL-based process executions. The BPEL output generated from the BPMN diagrams will also serve as the input source for this application.

2.1.1 Encoding norms in BPMN

After discussing the technical details behind BPMN and BPEL. We will demonstrate the effectiveness of BPMN as a tool for depicting a data compliance process in this project, below we will present some practical examples that used BPMN to achieve various process-based tasks.

2.1.1.1 Use case 1: eGovernment service

Palmirani and Governatori (2018) [7] demonstrated a use-case scenario related to an online academic platform which provides students and parents access to academic information and interaction with teachers, including chatline with specialized school staff. Art. 8 GDPR "*Conditions applicable to child's consent in relation to information society services 1. Where point (a) of Article 6(1) applies, in relation to the offer of information society services directly to a child, the processing of the personal data of a child shall be lawful where the child is at least 16 years old. Where the child is below the age of 16 years, such processing shall be lawful only if and to the extent that consent is given or authorised by the holder of parental responsibility over the child. Member States may provide by law for a lower age for those purposes provided that such lower age is not below 13 years.*" is applied to this case. Moreover, in order to access to the platform, the student must agree to general service condition and provide the consent for the controller's processing of personal data (Art. 4 GDPR), including sensitive data (Art. 6 GDPR). In their study on modelling legal knowledge for GDPR compliance checking, Palmirani and Governatori (2018) illustrate the process in Figure 2.1.

In this case, Palmirani and Governatori first analyze the tasks throughout the entire process and indicate the related GDPR articles of each tasks. The whole process utilizes four fundamental BPMN elements: the start event, which initiates the process; the end event, signifying the completion of the process; tasks, which represent the questions for the users, can involve a decision-making process or a confirmation of certain information; and exclusive gateways, which help decision-making by directing the process flow based on conditions.

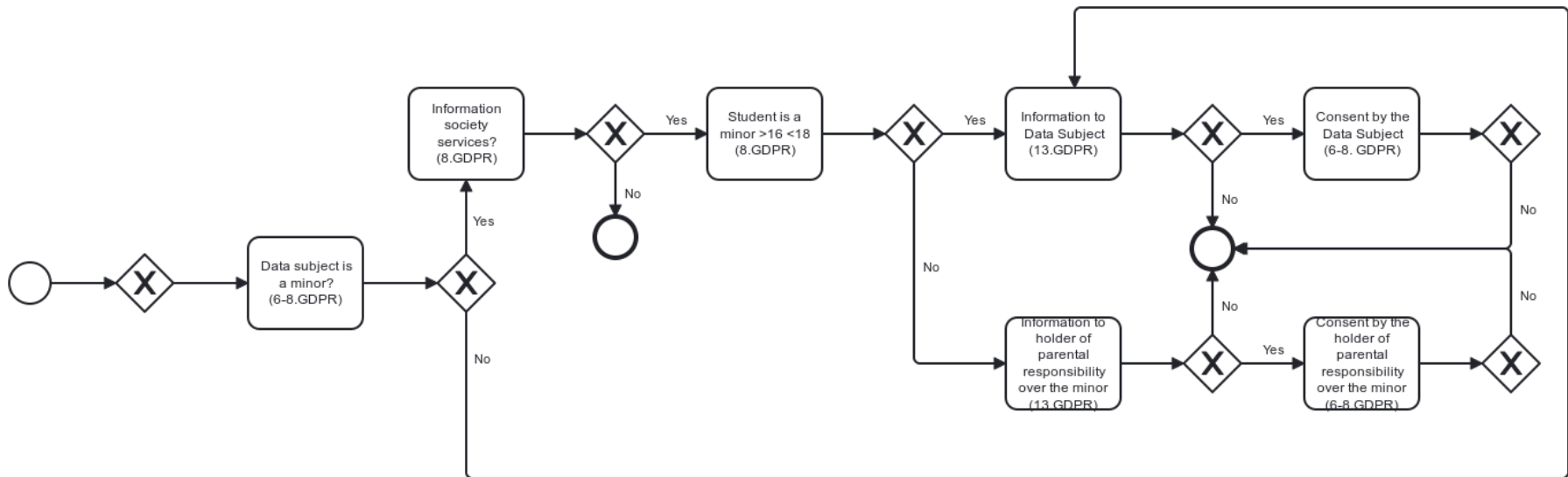


Figure 2.1: BPMN modelling of an eGovernment service (Palmirani & Governatori, 2018, p. 104)

2.1.1.2 Use case 2: Data Breach

In the study of S. Agostinelli et al (2019) [8], they focus on the obligations of the Data Controller, where the controllers need to adhere a list of constraints to be compliant with GDPR. These obligations include: *Data Breach, Consent to Use the Data, Right to Access and Rectify, Right of Portability, Right to Withdraw, and The Right to be Forgotten*. Figure 2.2 presents the constraint of Data Breach to showcase another example of modeling a legislative process with BPMN.

Figure 2.2 illustrates the procedure a Data Controller must follow upon encountering a data breach. The steps include identifying the affected Data Subjects and notifying the National Authority. The stolen data will be evaluated to determine if the Data Subjects need to be informed. It is important to note that these actions must be taken within 72 hours to avoid penalties of up to €20 million or 4% of global revenue.

In conclusion, BPMN provides a visual process management framework for handling data, which aids in identifying and managing potential privacy issues during data processing. In this example, a data breach response process simulated using BPMN can ensure that all data breach incidents are reported and handled according to the strict standards of the GDPR, reducing the risk of unlawful data processing.

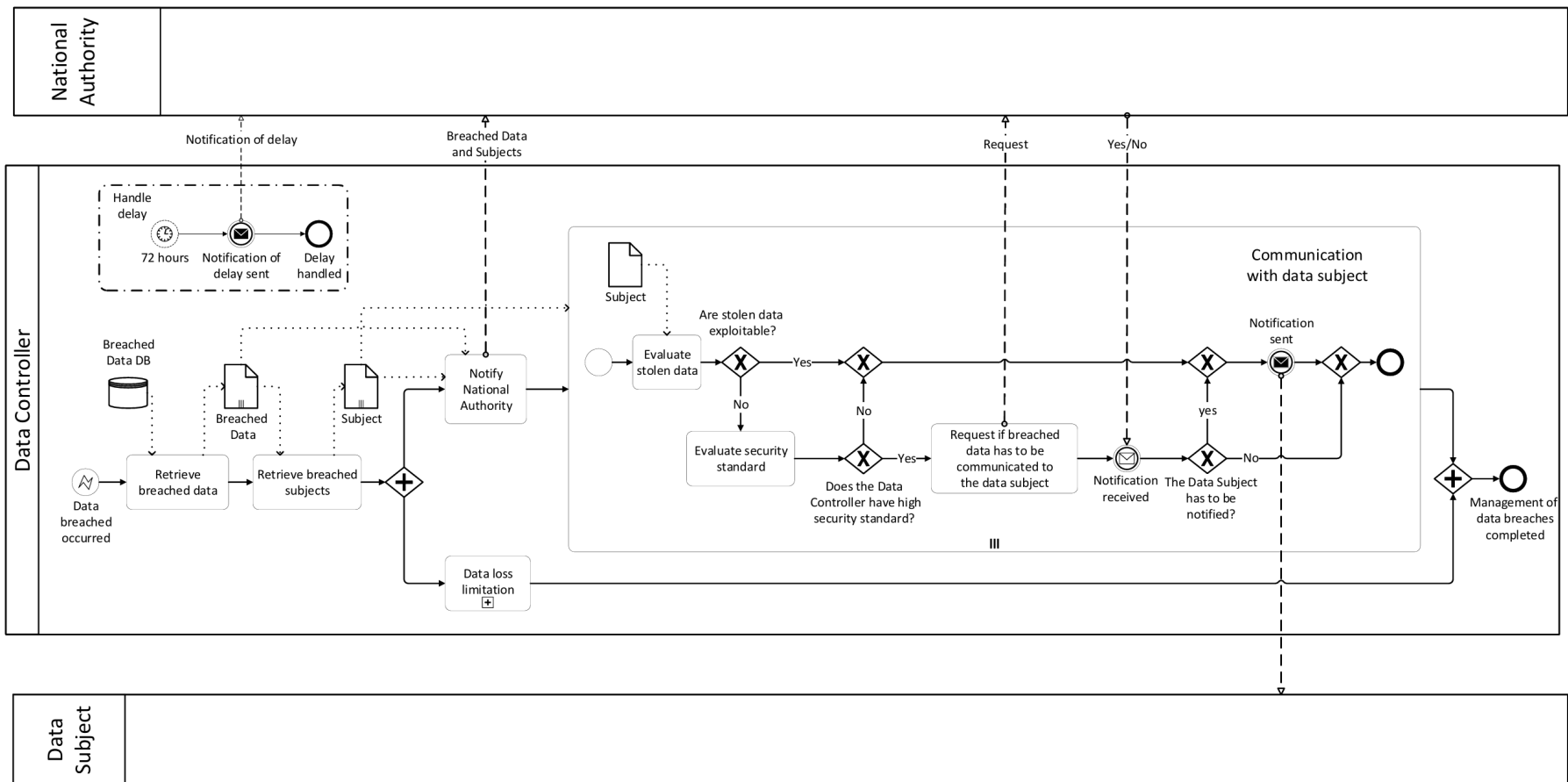


Figure 2.2: BPMN model for pattern Data Breach (S. Agostinelli et al, 2019)[9]

2.1.2 BPMN 2.0 diagram

With the elements and symbols that BPMN provides, it can accommodate processes ranging from simple to complex. This means that the complexity of the business scenario does not affect the usability of BPMN.

2.1.2.1 Elements and Symbols

This section will detail and explain the five elements in BPMN in order to represent any business process and essential symbols employed in this project, alongside clarifying their distinct roles within the BPMN framework, as presented in Table 2.1.

1. Flow objects: These include the most common elements within a diagram, including Events, Activities, and Gateways. They respectively represent what happens, the work being done, and the process flow based on conditions.
2. Connecting objects: These include Sequence Flows, Message Flows, and Associations. They respectively represent the order of activities, communication between different process participants, and the object links artifacts and data
3. Swimlanes: These include Pool and Lane to represent the participants in the process and the parts of the process for which they are responsible.
4. Artifacts: The purpose of Artifacts is to provide additional information to the process. These include data object, group and annotation.






Type	Symbol	Usage
Event	 Start Event	The process begins with the user being presented with the initial session.
Event	 End Event	The final process involves returning the classification result based on the specific conditions of each user.
Activity	 Task	This represents the process in which users clarify their situations during an interaction session.
Gateway	 Exclusive	This is the point at which the user makes a decision; a distinct choice will lead to a different result.
Gateway	 Inclusive	This is the point at which the user makes a decision between more than two options; it can be used for text-input answers or multiple-choice questions.

Table 2.1: Symbols and their usage

2.2 Tools for Regulation Assessment

Wizard-like webpages are a common style in today's web design because they can easily guide users through complex processes on the site. Simply put, a Wizard-like webpage is an interactive web interface that leads users through a series of steps in a linear fashion, making all operations easy to understand and execute, thereby increasing the completion rate. The key to designing a guided webpage is to provide clear instructions, progress feedback, and readability for the user, which effectively enhances the user experience. Moreover, the interactive features on the webpage are also key to making it easier for users to engage with the entire process. Based on the above, the design of guided webpages is also suitable for the objectives of this project. Below is an example of its practical application.

2.2.1 PET Explorer

Many companies and organizations possess a wealth of data beneficial to society, such as in health, poverty reduction, environmental protection, national security, and education. Linking these databases together can address some social issues and enhance public welfare. However, these databases often contain sensitive or private information, which cannot be freely shared with the public. Thus, Privacy Enhancing Technology (PET) has emerged. Under this technology, organizations can verify compliance with data-related regulations without providing complete data sets. The PET Explorer³ [10] provided by Sensor Lab at TNO offers a solution to explore Privacy Enhancing Technology (PET) in an organization and can be useful when performing a Data Protection Impact Assessment (DPIA). This tool also exemplifies the practical application of guiding users through a decision-making process, displaying the core principles of wizard-like design.

The decision tree interface provided by PET Explorer initially displays a starting question to the user upon entering the page, allowing them to choose an appropriate scenario. Based on the user's choice, a corresponding second question will appear below the initial one. After selecting an answer to the second question, a third question is shown, continuing until reaching the final node of the decision tree, which is the final decision result. What's special about this method is that it retains the user's decision-making process on the page, allowing them to modify their answers at any time and continue with a new process until the end. These designs are not only intuitive but also provide immediate feedback on the choices made by the user, all of which contribute to a good user experience.

2.2.2 High-Risk AI

As the application of artificial intelligence becomes increasingly widespread, understanding the potential impacts of an AI application is crucial. From an ethical perspective, a high-risk AI system could severely affect individual rights. Therefore, identifying whether an AI system is trustworthy can better facilitate risk management. Ensuring that the system complies with regulations can also enhance user and market trust in artificial intelligence. To tackle this issue, D. Golpayegani et al. (2023) [11] presented an application using semantic web technologies that helps

³<https://pet-explorer.sensorlab.tno.nl/decision-tree/>

State of the Art

users to easily assess whether their AI system is high risk or not based on the Annex III of the EU AI Act.⁴

This high-risk assessment application⁵ uses a single-page form that provides users with several drop-down lists to fill in features of the AI system, including the domain, purpose, capability, user, and entity. These five concepts, mentioned in Annex III, are required for determining high-risk AI. After submitting the answers, the application analyzes the combination of the user's answers and returns the assessment result. The system is considered safe if it identifies the AI as "likely to be Not High-Risk." However, if the combination meets any condition listed in Figure 2.3, the application will identify it as high risk and indicate which part of Annex III mentions such a condition. Additionally, it provides a list of the answers as a summary of the users' choices. This feature is useful to allow users to review their answers and ensure the correctness of the assessment.

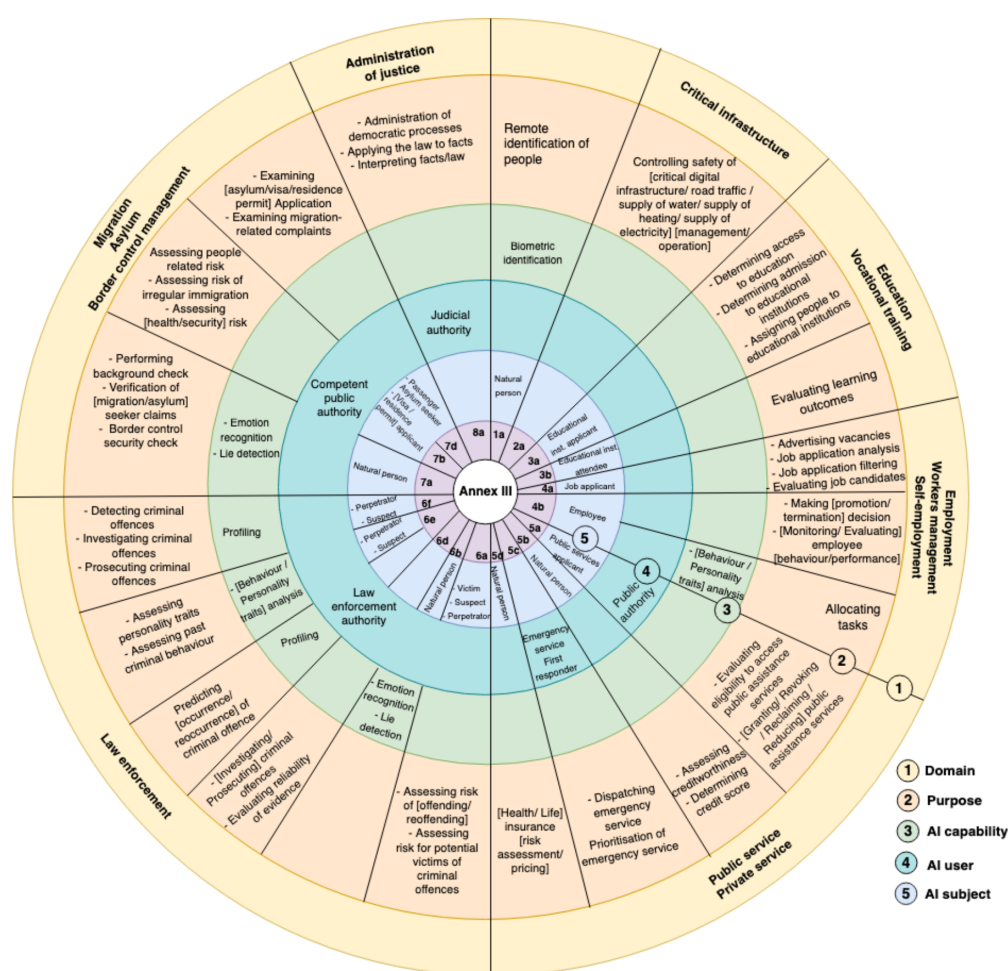


Figure 2.3: Describing Annex III high-risk conditions using the 5 concepts[11]

⁴<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>

⁵https://github.com/DelaramGlp/highrisk_app

2.2.3 Data Protection Self Assessment

To help small to medium-sized organizations assess their compliance with data protection law, the Information Commissioner's Office (ICO) presents Data protection self assessment tool⁶ to guide users in self-assessing their data compliance with categorized questions. This assessment tool covers both data controller and processor roles, which the user needs to identify on their own before selecting the corresponding checklist. After entering to the checklists, the user need to answer several questions related different part of the regulation. The answer options are categorize into four uniform options: "Not yet," "Partially," "Successfully," and "Not applicable." Each question also provides additional information to help users better understand the question and avoid misleading answers. When the answers are submitted, the system will then aggregate the scores from each response to generate an overall rating and suggest necessary actions to remedy any identified shortcomings along with guidance.

2.2.4 ValidaCripto RGPD

ValidaCripto RGPD⁷ is a tool presented by the Spanish agency of data protection for validating cryptographic systems, which are designed to protect personal data during processing. The "Guidelines for the Validation of Cryptographic Systems in Data Protection"⁸ document, based on GDPR requirements, provides a detailed description of the standards and procedures for effective encryption techniques. Most importantly, it evaluates various elements of the encryption system, including message space, key management, and encryption protocols, among others.

At the top-left of the page shown in Figure 2.4 is a horizontal slider, the user should select the level of criticality of the encryption using it. And enter data processing name in the text-box on the top-right.

⁶<https://ico.org.uk/for-organisations/advice-for-small-organisations/checklists/data-protection-self-assessment/>

⁷<https://validacriptorgpd.aepd.es/>

⁸<https://www.aepd.es/guides/guidelines-validation-cryptographic-systems-data-protection-processing.pdf>

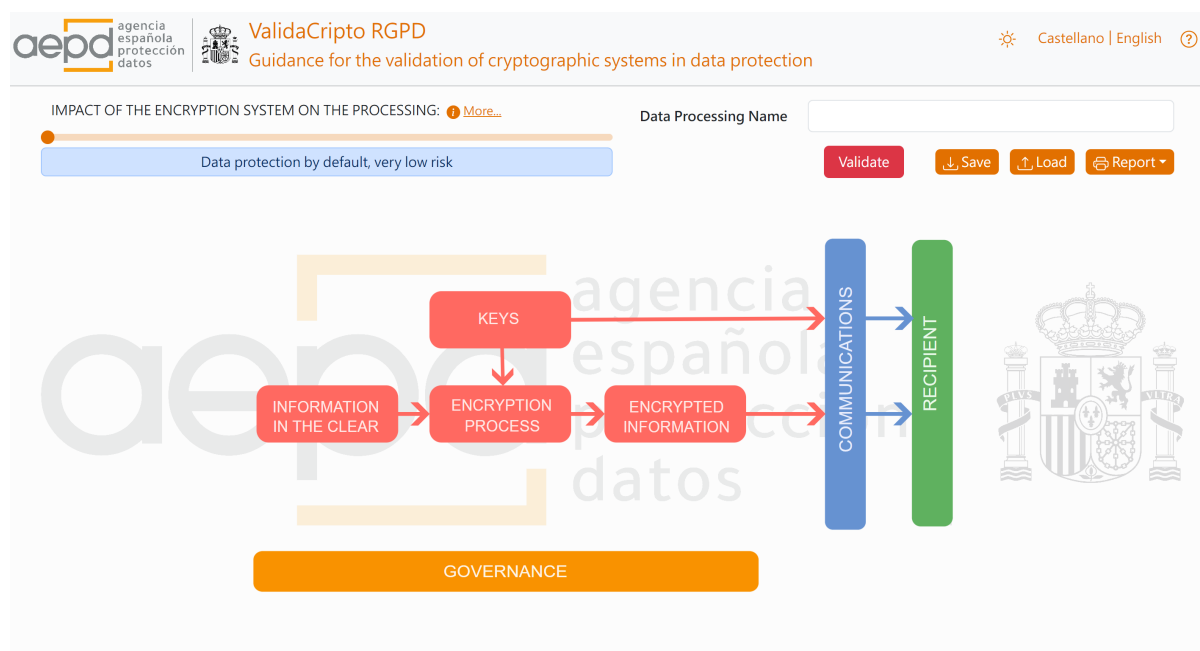


Figure 2.4: User interface of ValidaCripto RGPD

The page displays components of the encryption system in a diagram. Users need to hover their mouse over each component to reveal the elements underneath as shown in Figure 2.5. The small colored circle next to each item name represents the recommended validation level: optional (yellow), medium (orange), high (red).

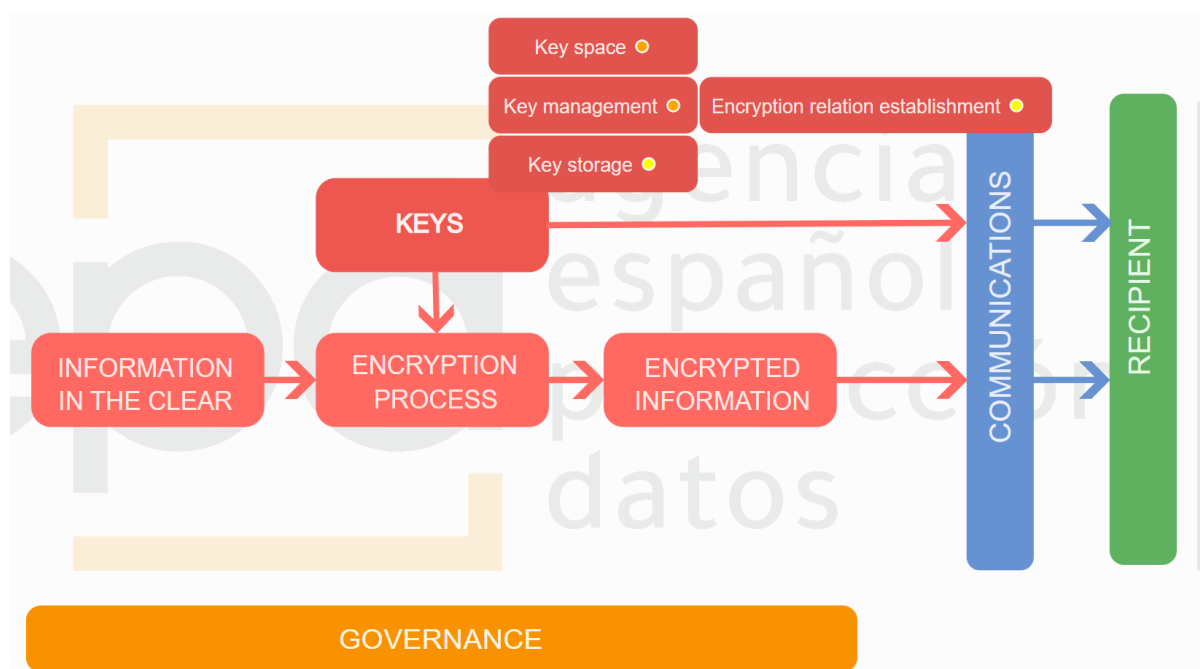


Figure 2.5: Elements under the component KEYS

Next, clicking on an element will display a table containing a brief description and various items that need to be validated, as shown in Figure 2.6. Users can select the corresponding status for individual items. The four statuses are: Unchecked, Not

2.2. Tools for Regulation Assessment

applicable, Fail, and Comply. Users can also add comments to each item by clicking the pencil icon. When the user has completed validating all the elements, they can click on the red Validate button at the top. A summary table will then be displayed showing the values recommended based on the criticality level previously selected and those obtained in the validation process.

Control	Validation
Key security requirements have to be defined, e.g. length over a number of bits (see Annex), format etc	Unchecked
There is a procedure to avoid reuse of keys.	Not applicable
It is not possible to use manually generated keys.	Fail
Passwords are not used as keys.	Comply

Acceptar

Figure 2.6: Validation section

Chapter 3

Analysis

In this chapter, we will provide a detailed introduction to the scenarios used on the website, the regulations behind these scenarios, and why automating this task is important. We will use BPMN diagrams to visualize the decision-making process, as well as present an XML summary that represents these diagrams.

Legal Compliance and Ethical Responsibility

This work does not pose any ethical issues. It does not process personal data or use copyrighted works. Furthermore, it does not create any legally relevant issues.

FAIR Principles

This research has adopted the FAIR principles[12]:

- **Findability:** The research data and results are stored in an open GitHub repository, available at Regulation-Assessment-for-Data-Spaces¹.
- **Accessibility:** The repository is open to everyone, allowing free access to the data.
- **Interoperability:** The work is licensed under the Apache License, Version 2.0, ensuring interoperability
- **Reuse:** The research data and results are available for reuse under the Apache License 2.0, details of which can be found at [apache.org](https://www.apache.org/licenses/LICENSE-2.0)².

3.1 Scenarios

3.1.1 Copyright Scenario

The scenario presented here uses copyright as an example. Copyright is a right immediately owned by the author once a work is completed, and it includes both economic and moral rights. These rights guarantee the author's control over their work and remuneration for its use through selling or licensing, as well as the rights to claim authorship and to refuse modifications of the work.[13] Others cannot publish or reproduce copyrighted works on behalf of the original author, so individuals must be cautious when citing someone else's work. We used six questions to determine

¹<https://github.com/chiachihsu/Regulation-Assessment-for-Data-Spaces>

²<https://www.apache.org/licenses/LICENSE-2.0>

whether the work can be used or if the user should contact the author to obtain permission for use. Below are the six questions included in the scenario shown in Figure 3.1.

1. **Is the work you intend to use one of these categories?**

Categories: Literary, dramatic, musical, poetry, novels, movies, songs, computer software, and architecture.

According to the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), Art. 2(2)³, *the expression “literary and artistic works” shall include every production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression.* Therefore, this question is designed to verify whether the work is protected by copyright law or if the user can use it directly.

2. **Does the author still alive?**

According to Directive (EU) 2006/116, Art. 1(1)⁴. *“The rights of an author of a literary or artistic work within the meaning of Article 2 of the Berne Convention shall run for the life of the author and for 70 years after his death, irrespective of the date when the work is lawfully made available to the public.”* Hence, this question is added to the process in order to verify the author’s status.

3. **Is your use covered by an exception?**

Exceptions: Education, quotation, parody, news reporting, research, private study, text & data mining, orphan work etc.

There are some exceptions to the copyright law mentioned in Directive (EU) 2019/790⁵ allow limited use of copyrighted work with copyright owner’s permission. Including but not limited to research, innovation, education and preservation of cultural heritage, text and data mining for the purposes of scientific research etc. The purpose of this question is to further confirm whether the use of the work falls under the limited exceptions of copyright protection after confirming that the author is still alive.

4. **Have more than 70 years passed since the death of the author (or the last surviving author)?**

As mentioned in Directive (EU) 2006/116, Art. 1(1), once the author has been deceased for over 70 years, the work automatically enters the public domain, meaning it can be freely used by anyone without needing permission from the author. For this reason, if the answer to question 2 is no, then the user will be asked to check how long it has been since the author passed away.

5. **Has the copyright owner licensed the work for public use?**

Since the use of licensed works varies depending on the nature of the license it possesses, this question verifies whether the work is licensed for public use or is limited by the restrictions of the license.

6. **Is your use covered by the license?**

The type of license can determine how others may use the work. Therefore, it is essential to check if the intended use is authorized by the work’s license.

To incorporate the decision-making process into the web application, we use BPMN to

³<https://www.wipo.int/wipolex/en/text/283698>

⁴<https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32006L0116>

⁵<https://eur-lex.europa.eu/eli/dir/2019/790/oj>

Analysis

generate the decision process. Figure 3.1 demonstrates the BPMN diagram generated based on the six questions mentioned in Section 3.1.1. Every question corresponds to a decision node, using arrows to represent the flow direction from one node to another. Each node has two options, pointing to different directions respectively. This method represents the basic functionality of the website, where the user can guide themselves through the process by clicking on the answer buttons. In this case, all the questions have only two options, making the operation intuitive and simple.

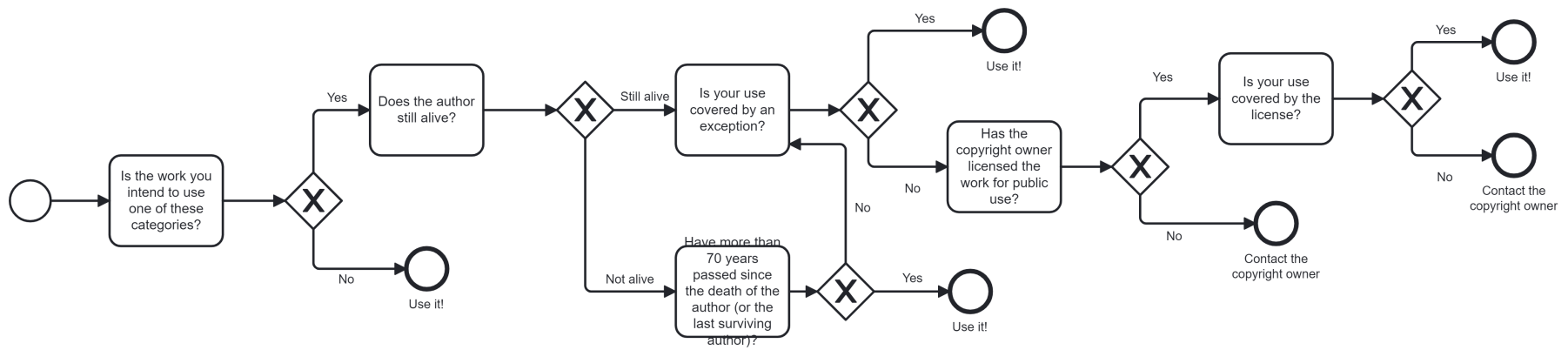


Figure 3.1: BPMN diagram of "Copyright" scenario

Analysis

The BPEL excerpt in Listing 3.1 illustrates an example of the process, which includes starting the assessment, checking if the work is protected by copyright, determining if the copyright holder is still alive, and directing the flow towards different decisions based on the evaluation results. Each step is connected by sequence flows, which guide the process through decisions that ultimately determine whether the work can be used legally. Every element has its own unique ID, and outgoing/incoming IDs where applicable. These IDs enable the application to link all the steps and decisions for dynamic interaction on the website.

Listing 3.1: BPEL Excerpt

```
1 <bpmn:process id="Process_0o9fbrm" isExecutable="false">
2   <bpmn:startEvent id="StartEvent_1dbwegp">
3     <bpmn:documentation>Assessment for ensuring that the use of creations complies with
4       copyright regulations.</bpmn:documentation>
5     <bpmn:outgoing>Flow_1lrf5u7</bpmn:outgoing>
6   </bpmn:startEvent>
7   <bpmn:task id="Activity_1no874j" name="Is the work you intend to use one of these
8     categories?">
9     <bpmn:documentation><b>Categories: Literary, dramatic, musical, poetry, novels, movies
10       , songs, computer software, and architecture.</b> Verify the category of the
11       creation, only certain types of works are protected by copyright.</bpmn:
12       documentation>
13     <bpmn:outgoing>Flow_102ntjb</bpmn:outgoing>
14   </bpmn:task>
15   <bpmn:exclusiveGateway id="Gateway_05f3eis">
16     <bpmn:outgoing>Flow_1jdk2p3</bpmn:outgoing>
17     <bpmn:outgoing>Flow_1ihulpe</bpmn:outgoing>
18   </bpmn:exclusiveGateway>
19   <bpmn:task id="Activity_0vsgn8x" name="Does the author still alive?">
20     <bpmn:documentation><p>In EU countries, copyright protects an author's intellectual
21       property until 70 years after their death or 70 years after the death of the last
22       surviving author in the case of a work of joint authorship.</p><a href="https://
23       europa.eu/youreurope/business/running-business/intellectual-property/copyright/
24       index_en.htm#inline-nav-1" target="_blank">More info</a></bpmn:documentation>
25     <bpmn:outgoing>Flow_143ponq</bpmn:outgoing>
26   </bpmn:task>
27   <bpmn:endEvent id="Event_0y3o764" name="Use it!">
28     <bpmn:incoming>Flow_1jdk2p3</bpmn:incoming>
29   </bpmn:endEvent>
30 </bpmn:process>
```

After uploading the XML files to the website, the website converts the XML files into scenario option buttons. Figure 3.2 displays the actual presentation of the website after the user selects a scenario they are interested in. The "Copyright" scenario is a decision-making process based on the process shown in Figure 3.1, where the user directs the process flow by clicking on answer buttons provided by the questions.

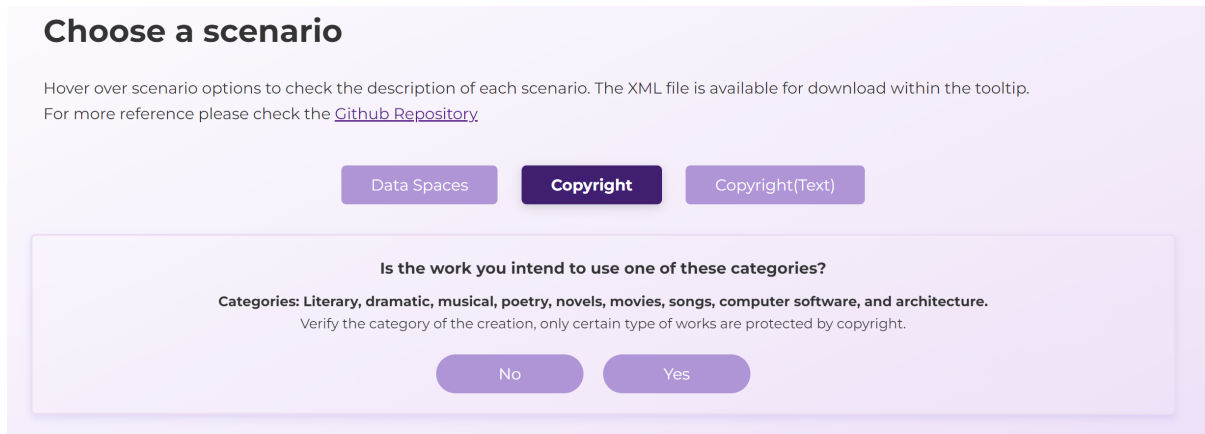


Figure 3.2: Demo of the question in "Copyright" scenario

3.1.1.1 Copyright Scenario - Method II

To enhance the extensibility of the website, we have generated the "Copyright(Text)" scenario, where the process flow is determined by the text input entered by the user. To fulfill this objective, the BPMN diagram needs to use an Inclusive Gateway to represent the possibilities of all potential answers instead of an Exclusive Gateway. Figure 3.3 shows an example of the same copyright scenario but with a text-input function for the first question of the process.

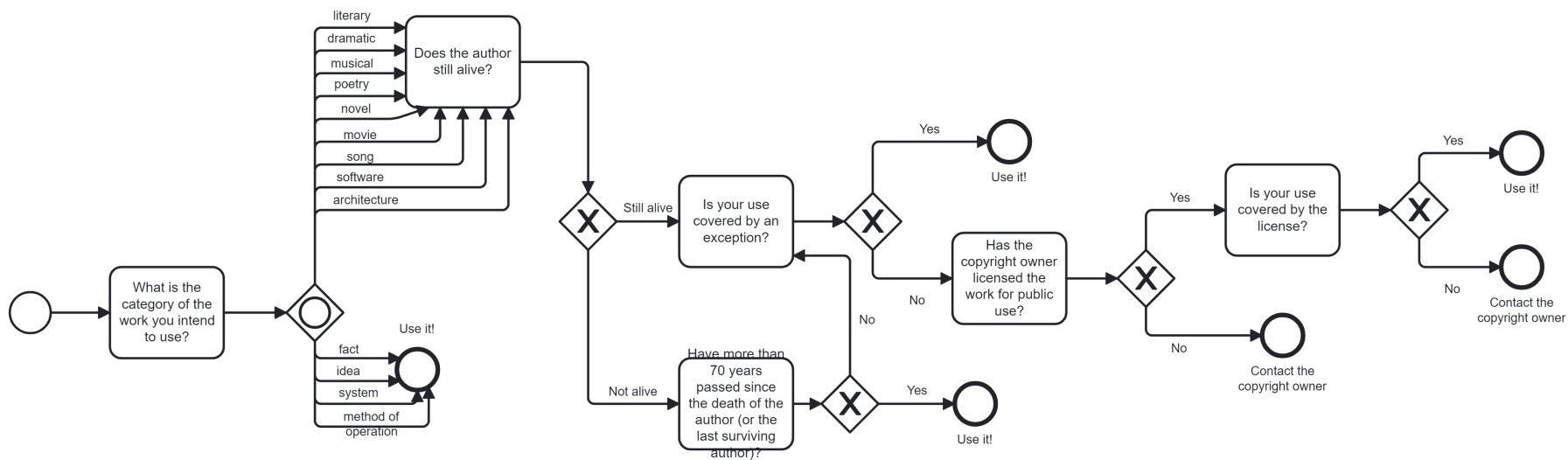


Figure 3.3: BPMN diagram of "Copyright_text" scenario

Figure 3.4 demonstrates the screen displayed when the user selects the "Copyright(Text)" scenario on the website. A text box is provided for entering the answer to this question. The system suggests potential answers in the description section under the question for error prevention purposes. The process design allows the user to enter the answer in either lower case or upper case, and subtle typos can also be processed. If the user's input does not match any potential answers included in the BPMN diagram, an error message "Please try again. No close match found for the input." will be returned, and the user can re-enter the answer. If a match is found between potential answers and the user's input, the process will move forward accordingly.

Choose a scenario

Hover over scenario options to check the description of each scenario. The XML file is available for download within the tooltip. For more reference please check the [Github Repository](#).

Data Spaces Copyright **Copyright(Text)**

What is the category of the work you intend to use?

Possible response: literary, dramatic, musical, poetry, novel, movie, song, software, architecture, fact, idea, system, method of operation

Verify the category of the creation, only certain type of works are protected by copyright.

software

Submit

Figure 3.4: Demo of the question in "Copyright_text" scenario

3.1.1.2 Data Spaces Scenario

This Data Spaces scenario aims to help non-legal experts conduct an initial assessment of whether their data is suitable for sharing in the Data Space. This process is divided into five levels: Initial Assessment, Proprietary Data Check, Data Usage Approval, Legal Compliance Check (Data Governance Act), and Compliance with the Data Act.

Level 1: Initial Assessment

The first level is a preliminary check to ensure that the participation is related to accessing or manipulating data in Data Space. If it is not involved, then it is not necessary to continue with the compliance-checking process.

- Does your participation in the Data Space involve accessing or manipulating data?

Level 2: Proprietary Data Check

The second level checks the propriety and confidentiality of the data. If the data is proprietary or confidential, it is required to assess whether the data is necessary for the decision according to Lawfulness of processing in GDPR Art. 6⁶. If it is not, then the user should consider other data sources.

⁶<https://gdpr-info.eu/art-6-gdpr/>

- Is the data considered proprietary or confidential?
- Is the data necessary for the decision?

Level 3: Data Usage Approval

The third level ensures the data usage complies with established policies. If data usage policies exist, the proposed usage method needs to be checked for compliance with these policies. If there are no policies, relevant authorization from stakeholders needs to be sought. This level corresponds with GDPR Art. 24 Responsibility of the controller ⁷, which requires that *“the controller shall implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation.”*

- Is there a data usage policy in place for the Data Space?
- Does the proposed usage align with the policy?

Level 4: Legal Compliance Check (Data Governance Act)

To be consistent with the Data Governance Act, it is important to ensure the legitimacy of personal data processing. The fourth level checks whether the decision involves processing personal data. If personal data is involved, it is mandatory to obtain consent or have another legal basis for processing. If personal data is not involved, the user can proceed to the fifth level to check compliance with the Data Act.

*Legal basis: Consent, Necessary for the purpose of a contract, Compliance with law, Vital interests, Public interest, Legitimate interests. ⁸

- Does the decision involve processing personal data?
- Has the necessary consent or any other legal basis been obtained for processing?

Level 5: Compliance with Data Act

The fifth level aims to examine whether the data sharing involves any transfer of personal data which are undergoing processing or are intended for processing after transfer to a third country or to an international organisation.⁹ If it does, then a lawful basis is required to comply with the Data Act. If a lawful basis does not exist, it is recommended to seek legal advice or alternative methods.

- Does the data sharing in the data Space involve cross-border data transfers?
- Is there a lawful basis for transferring data outside the EU/EEA?

Figure 3.5 is the BPMN diagram generated base on the five-level assessment for Data space data sharing.

⁷<https://gdpr-info.eu/art-24-gdpr/>

⁸<https://www.hrb.ie/funding/gdpr-guidance-for-researchers/gdpr-overview/what-is-a-legal-basis/>

⁹<https://gdpr-info.eu/art-44-gdpr/>

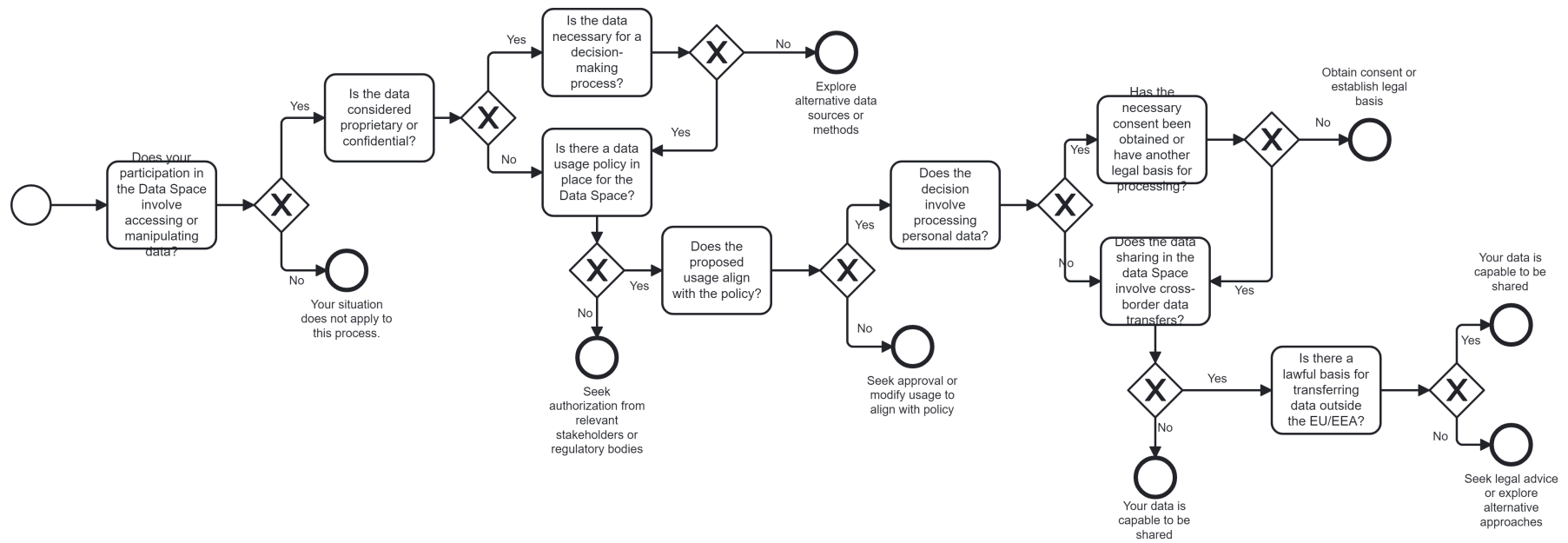


Figure 3.5: BPMN diagram of "Data Spaces" scenario

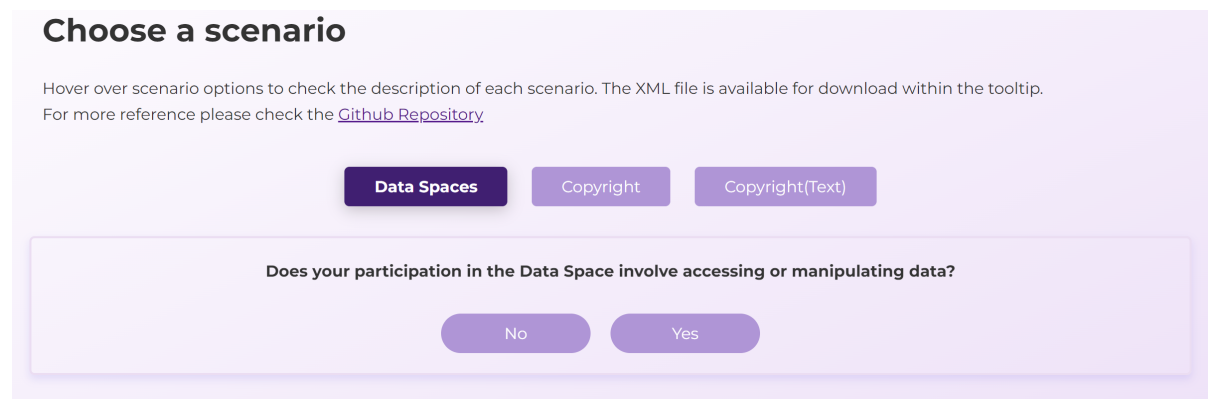


Figure 3.6: Demo of the question in "Data Spaces" scenario

3.2 Requirements

To achieve the objective of this project, a website application named "CompliancePal" is developed to assess data compliance with relevant EU regulations. This application has following requirement:

- R1. The application must be able to read and parse BPEL file from a designated source.
- R2. The application should be able to dynamically generate and display questions based on the BPMN elements parsed from the BPEL file. These BPMN elements include Start Event, Task, Exclusive Gateway, Inclusive Gateway, and End Event.
- R3. The application should keep the decision-making process visible to the user; this enables the possibility of reviewing the process and ensuring the correctness of the assessment result.
- R4. The application must provide functions for users to change their answers, including the ability to remove or add subsequent questions based on these changes.
- R5. The application should be capable of handling text-input questions and adjusting the decision-making flow based on the text input content.
- R6. The application should allow users to download the XML file for the provided scenarios.
- R7. The application needs to achieve the Technology Readiness Level of 3 (TRL3) where the technological concept should be validated through experimental proof-of-concept.[14]

Chapter 4

Results and conclusions

4.1 Implementation

In this project, we developed a method for verifying data compliance and integrated it into the wizard-like webpage called CompliancePal. This tool achieves a complete process from XML document parsing to user interaction through collaboration between the front end and the back end. The front-end design allows users to select the appropriate scenario from the buttons displayed on the page, and then presents the corresponding questions and options. The back end is responsible for providing BPEL files. Finally, this chapter evaluates CompliancePal to verify whether it meets the previously established requirements.

4.1.1 Architecture

The programming languages used in the website are HTML, CSS, JavaScript for the frontend and Flask for the backend. From a high-level perspective, the application will first load the module chosen by the user, and the system will parse the module and visualize it on the website as questions and options. During the process, several events need to be handled behind the scenes. For instance, the next question to be displayed is based on the user's answer, changes to the route in the decision-making process if the user changes the answer to a previous question, and the method to match the user's text input with potential answers, etc. Eventually, the assessment result along with suggested actions will be returned.

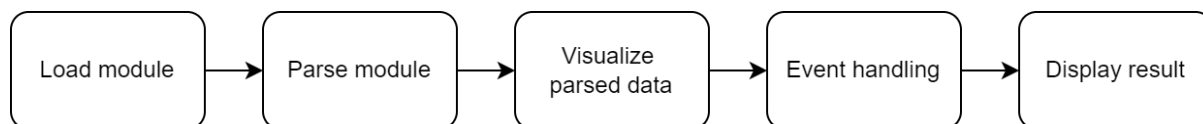


Figure 4.1: High-level system diagram

4.1.1.1 Frontend

The frontend design is primarily a wizard-like webpage, where users can select the appropriate scenario from option buttons based on their needs. After selection, question text and option buttons will appear below, and the user can choose the answer

Results and conclusions

that best describes their situation. If there are more than three options, it will be considered a short-answer question. The webpage will determine the next question to appear or the assessment result based on the user's input.

The logic behind the webpage is all contained within JavaScript code, including parsing XML files and displaying them on the webpage. Below, the logic of the script will be explained in six parts.

- **Initial Setup and Event Listener**

- *DOMContentLoaded* Event: The entire script is wrapped in an event listener that waits for the *DOMContentLoaded* event. This ensures that the script runs only after the entire HTML document has been fully loaded, making sure that all elements are accessible.

- **XML File Selection and Data Processing**

- Button Initialization: Each button is associated with an XML file. An event listener is set to handle click events that trigger the XML display function, and mouseover events that show tooltips.
- Tooltip Functionality: Tooltips display the description of the scenario fetched from the XML file, and a download button is provided for users to download the XML file.
- *onXmlFileSelect* Function: This function defines the next action when a item is selected by the bottom. It retrieves the selected file name and triggers the *displayXmlData* function with the selected file name as its argument.

- **Fetching and Parsing XML Data**

- *displayXmlData* Function: This function is designed to fetch XML data based on the selected file name passed from the *onXmlFileSelect* function. It uses the Fetch API to get the XML file, then parses the file using 'DOM-Parser' to convert the string into a document object model (DOM) that can be traversed and manipulated.

- **Processing BPMN Elements**

- The script then extracts and processes the BPMN elements that are used in this project: start event, tasks, exclusive gateways, inclusive gateways, sequence flows, and end events. The script will extract the IDs, names, incoming and outgoing flows of these elements and organizes this information into data structures *dataToDisplay*, which is an object mapping element IDs to their data, and *elementsByFlow*, an array of data objects for further processing.

- **Display Logic and Interaction**

- First Question Detection: To find out the first question of the BPMN diagram, we will search the *dataToDisplay* data structure to detect which incoming ID of a task is the same as the outgoing ID of the start event and save it in a variable *questionFlow*. This variable will then be passed to the *displayCurrentQuestion* function as its argument.
- Dynamic Interaction: This section consists of several functions to achieve

dynamic interaction, including displaying questions and options, and navigating through a set of questions based on the user's choices.

1. *displayCurrentQuestion* Function: For each parsed element that represents a potential question or decision point, *displayCurrentQuestion* function is called to create a visible question block within the UI.
2. *updateOptions* Function: If the next question has more than two options, it treats it as a short-answer question and returns a text box; otherwise, *createOptionButton* function will be triggered and two option buttons will be provided.
3. *createOptionButton* Function: This function creates a button element and sets its text to the name of the *optionElement*, along with adding an event listener. When clicked, the system saves the ID of the answer in *answerID*, highlights the selected option, and triggers the *updateFollowingQuestions* function to create the next question block. Additionally, if the user changes the answer to a previous question, this function detects the change, removes the subsequent questions, and updates the following questions accordingly.
4. *removeAllFollowingQuestions* Function: This function manages the dynamic update of questions when a user changes their answer to a previous question, affecting the subsequent flow of questions. It finds the index of the parent div of the current question and then enters a loop to remove any question with an index number greater than the current one from the display.
5. *resultBackgroundColor* Every click on the option button triggers this function, which includes a loop to iterate through all the questions on display and a conditional clause to detect whether the final assessment result has been returned. If so, the background color changes to a contrasting color to highlight the completion of the process.
6. *levenshteinDistance* Function [15]: This function calculates and returns the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one string into the other. It uses a dynamic method, building a matrix track that keeps track of distances between all possible substrings of *s* and *t*, representing user input and potential options, respectively. Levenshtein distance is applied for text input, with the threshold set at 3. This allows users to make minor typos when answering the question and displays the next question if a close match is found.

4.1.1.2 Backend

The backend is a Flask web application that serves the BPEL file of the BPMN process. This Flask application is designed to serve as part of a web application which focus on a section named "rads". It provides functionality to display an index page, a page listing BPEL files for assessments, and serves the actual BPEL files from a specified directory. It uses Python's `os` module to define a path to the directory (BPMN_FOLDER) where BPEL files are stored.

4.2 Evaluation

This section helps to verify whether the requirements mentioned in Section 4.1.1 have been satisfied and how they were fulfilled in CompliancePal.

R1. The application must be able to read and parse XML file from a designated source.

CompliancePal can load and parse any BPEL files that contain the elements mentioned in Table 2.1. This is achieved by clicking on the scenario buttons in the Assessment page.

R2. Dynamically generate and display questions based on the BPMN elements parsed from the XML file.

Since each BPMN element has its own incoming and outgoing ID, CompliancePal can easily determine which question to display by matching the incoming and outgoing IDs of each element. For example, the outgoing ID of Task A is identical to the incoming ID of Task B, and the outgoing ID of Task B would be the incoming ID of Task C, and so on. By connecting these IDs, CompliancePal can dynamically generate and display the questions until it reaches the end of the process.

R3. Keep the decision-making process visible to the user; this enables the possibility of reviewing the process and ensuring the correctness of the assessment result.

CompliancePal satisfies this requirement by adding a container for each question whenever a new question is generated, allowing all questions to remain on the screen along with the final result. Additionally, CompliancePal highlights the user's answers to each question to facilitate an easier review of the decision-making process.

R4. Provide functions for users to change their answers, including the ability to remove or add subsequent questions based on these changes.

CompliancePal satisfies this requirement by assigning a unique index to each question to locate it within the parent container. Whenever a click event on an answer is detected, the index is used to identify the question currently being answered. By implementing a loop check, all questions with a higher index number are removed. The updated following question or result is then displayed on the screen.

R5. Be capable of handling text-input questions and altering the decision-making flow based on the text input.

CompliancePal is able to handle text input by calculating the Levenshtein distance between the user's input and the built-in answers in the BPEL file.

R6. Allow users to download the XML file for the provided scenarios.

CompliancePal provides a download button for each XML file within the tooltip.

The GitHub repository behind the website is also provided in the footer on both the home page and the assessment page.

4.3 Contributions

- A novel method for transforming BPMN processes into web-based wizards.
- A practical, open-source tool that simplifies legal compliance tasks, highlighting the ease of adding new scenarios.
- An extensibility mechanism enables the tool to parse text input, demonstrating the possibility of incorporating complex evaluations.

4.4 Conclusions

During this project, I developed a website to simplify regulation compliance tasks. Different scenarios were tested, including copyright and data space regulations. This work is grounded in the idea that a preliminary compliance check can be made by non-experts in legal matters. Indeed, in non-trivial cases, the compliance concerns will be forwarded to legal departments, but laymen can act as a first layer, serving as an initial filter for legal issues, thus saving valuable resources in an organization. The project demonstrated that agile development methodologies and user-friendly process-design tools can efficiently produce compliance-support tools.

Standards play an important role in the generalization of technical solutions. Aware of this, the work relies on well-established industry standards such as BPMN. Choosing a standard-based solution has paid off for two reasons: Firstly, tools like Camunda facilitate automated workflow execution. Secondly, the universality of the language: the number of BPMN-literate users is high, which ensures a broad understanding among users, allowing them to grasp the whole intent of a legal provision from the mere representation of the process diagram.

4.5 Future Work

This study successfully developed a method to simplify regulation compliance tasks in data spaces. Future work could involve usability testing conducted by laymen to ensure that any user can easily understand how to use the website without additional instruction from the developer, expanding the functionality of this tool to include support for new data regulations, and improving the user experience and automation of the process. Additionally, incorporating machine learning techniques to further enhance the accuracy and intelligence of the tool is also expected.

Bibliography

- [1] Hamilton, J. (2023, January 30). Government rules as code – a transformative idea. <https://www.button.is/post/government-rules-as-code-a-transformative-idea>
- [2] Common European Data Spaces. (2024, March 13). Shaping Europe's Digital Future. <https://digital-strategy.ec.europa.eu/en/policies/data-spaces>
- [3] Otto, B., ten Hompel, M., & Wrobel, S. (2022). Designing Data Spaces: The Ecosystem Approach to Competitive Advantage (p. 580). Springer Nature.
- [4] Data Governance Act explained. (2024, January 25). Shaping Europe's Digital Future. <https://digital-strategy.ec.europa.eu/en/policies/data-governance-act-explained>
- [5] Education, I. C. (2022, January 6). The Basics of Business Process Modeling and Notation (BPMN) - IBM blog. IBM Blog. <https://www.ibm.com/blog/bpmn/>
- [6] Khalaf, R., Keller, A., & Leymann, F. (2006). Business processes for Web Services: Principles and applications. *IBM Syst. J.*, 45, 425-446. <https://doi.org/10.1147/sj.452.0425>.
- [7] Palmirani, M., & Governatori, G. (2018, December). Modelling Legal Knowledge for GDPR Compliance Checking. In *JURIX* (Vol. 313, pp. 101-110).
- [8] Agostinelli, S., Maggi, F. M., Marrella, A., & Sapio, F. (2019). Achieving GDPR compliance of BPMN process models. In *Information Systems Engineering in Responsible Information Systems: CAiSE Forum 2019, Rome, Italy, June 3–7, 2019, Proceedings 31* (pp. 10-22). Springer International Publishing.
- [9] Fig. 2. BPMN model for pattern Data Breach. (2019, May). ResearchGate. https://www.researchgate.net/figure/BPMN-model-for-pattern-Data-Breach_fig2_333312868
- [10] CBS, KNB, Rabobank, University of Maastricht, & TNO. (2024). TNO PET explorer - PET Decision Tree. <https://pet-explorer.sensorlab.tno.nl/decision-tree/>
- [11] Golpayegani, D., Pandit, H. J., & Lewis, D. (2023, June). To Be High-Risk, or Not To Be—Semantic Specifications and Implications of the AI Act's High-Risk AI Applications and Harmonised Standards. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 905-915).

- [12] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.
- [13] Copyright in the EU: How to get copyright protection - Your Europe. (2023, August 11). Your Europe. https://europa.eu/youreurope/business/running-business/intellectual-property/copyright/index_en.htm#inline-nav-1
- [14] Wikipedia contributors. (2023, December 11). Technology readiness level. Wikipedia. https://en.wikipedia.org/wiki/Technology_readiness_level
- [15] Wikipedia contributors. (2024, February 23). Levenshtein distance. Wikipedia. https://en.wikipedia.org/wiki/Levenshtein_distance

Annex I - Github README.md

This README markdown file is in the GitHub repository of this work. It includes a brief introduction to the work, an example of a regulation assessment scenario, instructions on how to run the application locally, acknowledgments, and a disclaimer. By reading this file, the user can get an overview of the work and have quick access to the results.

Regulation Assessment for Data Spaces

The aim of the project is to develop a tool to ease data compliance tasks. In the framework of Common European Data Spaces, data needs to be assessed in terms of its compliance with new relevant EU regulations and policies. The tool would provide comprehensive reports and recommendations to help organizations ensure that their data spaces meet the necessary regulatory standards.

The objective of the system is to guide the user towards discovering the legal obligations derived from the data/AI operations made. The system has the form of a wizard-like tool, where the user answers questions in a series of steps, and gets a report in the last step depending on the given answers. We claim that a general tool is possible, where the logic is detached from the actual questions and results. In order to make this tool general, a decision tree parser will be programmed. The logic of the system may be encoded in BPMN (flow of the decision tree), which can be encoded as XML (BPEL).

Figure 1: Introduction of the project

Example: Regulation assessment checklist for Data Spaces data sharing.

This example demonstrates an assessment process and how to operate the app.

Diagram of the process

This assessment process includes Proprietary Data Check, Data Usage Approval, Legal Compliance Check of Digital Governance Act (DGA) and Data Act

```
graph TD
    Start(( )) --> Q1{Does your participation in the Data Space involve accessing or manipulating user?}
    Q1 -- No --> End1((Your situation does not apply to this process.))
    Q1 -- Yes --> Q2{Is the data considered proprietary or confidential?}
    Q2 -- No --> End2(( ))
    Q2 -- Yes --> Q3{Is there a data usage policy in place for the Data Space?}
    Q3 -- No --> End3(( ))
    Q3 -- Yes --> Q4{Is the data necessary for a decision-making process?}
    Q4 -- No --> End4((Explores alternative data sources or methods))
    Q4 -- Yes --> Q5{Does the proposed usage align with the policy?}
    Q5 -- No --> End5((Seek authorization from relevant stakeholders or regulatory bodies))
    Q5 -- Yes --> Q6{Does the decision involve processing personal data?}
    Q6 -- No --> End6(( ))
    Q6 -- Yes --> Q7{Does the data sharing in the data Space involve cross-border data transfers?}
    Q7 -- No --> End7(( ))
    Q7 -- Yes --> Q8{Has the necessary consent been obtained or have another legal basis for processing?}
    Q8 -- No --> End8((Obtain consent or establish legal basis))
    Q8 -- Yes --> Q9{Is there a lawful basis for transferring data outside the EU/EEA?}
    Q9 -- No --> End9((Your data is capable to be shared))
    Q9 -- Yes --> End10{Seek legal advice or explore alternative approaches}
    End10 --> End10
```

Figure 2: Example of the application

Run the app locally

Step 1: Clone the Github repository

```
git clone https://github.com/chiachihsu/Regulation-Assessment-for-Data-Spaces.git
```

Step 2: Initialize the Environment

Install the necessary dependencies using the following command:

```
pip install -r requirements.txt
```

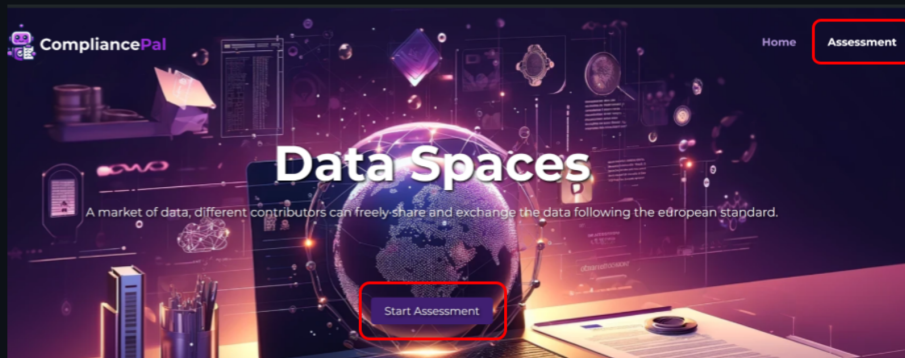
Step 3: Run the application

Start the application with following command:

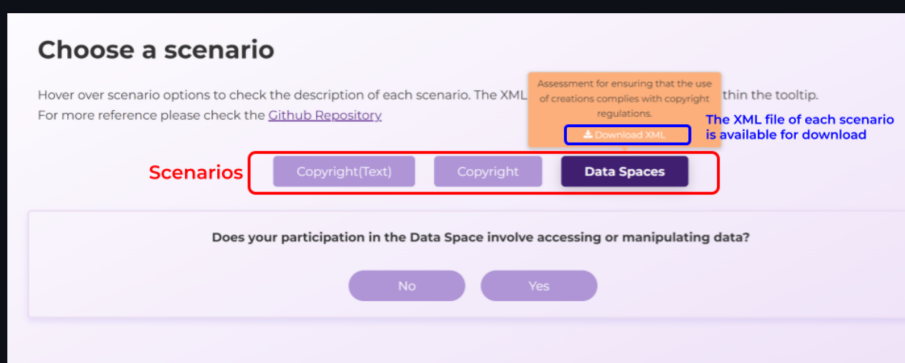
```
python app.py
```

Step 4: Use the application

1. Open a browser and navigate to <http://127.0.0.1:7000>
2. Click the 'Start Assessment' button in the middle of homepage or the 'Assesmet' tab on the top-right corner.



3. Choose a scenario.



4. Start the assessment process.

Possible extension

The users can modify/create XML file and place the file in the 'BPMN' folder to add a new process into the app.

Figure 3: App Instruction (Local)

Run the app online

Assess to the app directly through the following link:

<https://reg-assess-data-spaces-ac4fc5063977.herokuapp.com/>

Figure 4: App Instruction (Online)

Acknowledgements

INESData (<https://inesdata-project.eu/>) is a Spanish Incubator of Data Spaces and AI Services using federated infrastructures in the Cloud. It focuses on simplifying the tech adoption and accelerating the industry deployment of the national Data Space ecosystem by contributing with four Data Spaces (language, mobility, legal and public tender, and media) to demonstrate the benefits of Data Spaces and applicability of the related technology. It is funded by the Spanish Ministry of Digital Transformation and NextGenerationEU, in the framework of the UNICO I+D CLOUD Program - Real Decreto 959/2022.

Disclaimer

This work is for research purposes, and you should not trust the answers.

Figure 5: Acknowledgements and Disclaimer