Dr.Aid: supporting data-governance rule compliance for decentralized collaboration in an automated way

ANONYMOUS AUTHOR(S)

Collaboration across institutional boundaries is widespread and increasing today. It depends on federations sharing data that often have governance rules or external regulations restricting their use. However, the handling of data governance rules (aka. data-use policies) remains manual, time-consuming and error-prone, limiting the rate at which collaborations can form and respond to challenges and opportunities, inhibiting citizen science and reducing data providers' trust in compliance. Using an automated system to facilitate compliance handling reduces substantially the time needed for such non-mission work, thereby accelerating collaboration and improving productivity. We present a framework, Dr.Aid, that helps individuals, organisations and federations comply with data rules, using automation to track which rules are applicable as data is passed between processes and as derived data is generated. It encodes data-governance rules using a formal language and performs reasoning on multi-input-multi-output data-flow graphs in decentralised contexts. We test its power and utility by working with users performing cyclone tracking and earthquake modelling to support mitigation and emergency response. We query standard provenance traces to detach Dr.Aid from details of the tools and systems they are using, as these inevitably vary across members of a federation and through time. We evaluate the model by encoding real-life data-use policies from diverse fields, showing its capability. We argue that this approach will lead to more agile, more productive and more trustworthy collaborations and show that the approach can be adopted incrementally. This, in-turn, will allow more appropriate data policies to emerge opening up new forms of collaboration.

CCS Concepts: • **Security and privacy** → **Usability in security and privacy**; • **Social and professional topics** → **Computing / technology policy**.

Additional Key Words and Phrases: cross-boundary collaboration, data policy, governance rules, formal methods

ACM Reference Format:
Anonymous Author(s). 2018. Dr.Aid: supporting data-governance rule compliance for decentralized collaboration in an automated way. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 35 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The collaboration across institutional boundaries is an increasing practice in scientific research today, whether through tight research alliances or loosely coupled research federations. In the collaboration, data sharing is one of the core activities involved, as well as computation. There are initiatives such as (linked) open data [63], Research Objects [27] or FAIR [64] to provoke data sharing to wider audiences, to improve reproducibility of science, to broaden impact, etc. However, in many cases, data providers or governors need to establish and extend data governance rules, due to governmental policies or the properties of the data itself (e.g. containing sensitive information) [45]. In such circumstances, it is intrinsically impossible to simply make it "open data". Current practice for such situations often requires data users to submit applications and undergo training on security, privacy, sensitivity[1] and ethical data management before gaining access to the data, and their results may also require to be screened before they are allowed to disclose them to a wider audience. This is a tedious and time consuming procedure that deters researchers from using data, even if the data governance rules only pertain to a small portion of the data.

This issue is a socio-technical problem in three aspects: its origin, its impact, and the way to solve it. The problem originates from the societal requirements that data governance is necessary (e.g. to acknowledge the data's authors, to reduce unexpected harm, etc) [32, 35]. When data is used and processed on computers other issues arise (e.g. copying and transmitting data is almost cost-free, computer systems often do not track data lineage). This creates a polarization of data governance practice in real-life: data are either completely open with unenforced licensing rules, or under strict protection. This slows down the research progress, requiring more manpower for non-mission work, and limits reproducibility. Researchers using data may also be required to work in restrictive environments or limited in the technologies they may use [58]. This situation may be exacerbated for research taking data from different sources – the union of the prevailing rules may be hard to fathom out and even make the work incompatible with full compliance. Therefore, overcoming this problem requires improving the technology we have, to enable systematic monitoring and enforcement of data use policies, while gradually shifting the social practice. In the end, a new paradigm of computer supported rule formulation and compliance will emerge to facilitate collaborative work.

Taking a broader view, this issue applies beyond traditional research data and forms. It covers the technologies and methods to promote reproducibility in non-traditional data, e.g. social media [35], the discussion about the issues in traditional consent-based user agreement [41, 50], and emerging issues for IoT (Internet of Things) or smart devices [21, 60, 66]. In particular, they all pose the challenges with non-centralized data processing. Different approaches try to tackle the issue with different viewpoints and goals (Section 2). Our approach is to model data-governance rules in a computer-interpretable form, and to use (retrospective or proactive) data-flow-trace information to (semi-)automatically assist with compliance.

Although different models have different features and focuses, there are two major reasons for using a formal model: (1) to avoid the ambiguity in natural languages; (2) to expose/extract the similarity in data-governance rules, despite their representational heterogeneity. Figure 1 presents rules on selected data-governance rules available online (usually under the name of Terms of Use), and highlights the most informative parts It can be noticed that most parts are less informative or even unimportant to the policies themselves. Using a formal model can reduce such issues, and make the rules concise and accurate.

In our research, we take account of the data-governance context: data are from different sources and are processed by different bodies; the data processors are in different institutions who may not have tight collaboration agreements. Output data can be taken as input for other work, immediately as part of a current campaign, or in a currently unplanned future campaign involving different partners. We call this a *federated data-processing context.* Such a context is aligned with data-intensive research [34], where data have a wide-range of different governance requirements (policies). Collaboration across institutional boundaries is common practice for such a context. It is essential to properly comply with the data-governance rules, otherwise a collaboration may collapse and future collaborations with the same partners may become unachievable.

To set the scene, we present an example extracted from research practice: *Dataset (D) comes from a data provider (DP). It contains some sensitive information in its column "*`DoB`*" (Date of Birth). The rest of the data is not sensitive. DP wants to*

[1]Sensitive encompasses personal data, commercial-in-confidence and content such as emergency-response locations to avoid panic and media.

I agree to restrict my use of CORDEX model output for non-commercial research and educational purposes only. [1]

In publications that rely on the CORDEX model output, I will appropriately credit the data providers by an acknowledgement similar to the following: "We acknowledge..." [1]

You may extract, download, and make copies of the data contained in the Datasets, and you may share that data with third parties according to these terms of use. [2]

When sharing or facilitating access to the Datasets, you agree to include the same acknowledgment requirement in any sub-licenses of the data that you grant, and a requirement that any sub-licensees do the same. [2]

Data is non-transferrable (other than as permitted in the licence) and confidential in nature. [3]

Data is not to be used to identify, contact or target patients or general medical practitioners. [3]

[1] CORDEX terms of use: https://www.hereon.de/imperia/md/assets/clm/cordex_terms_of_use.pdf

[2] World Bank Terms of Use for Datasets: https://www.worldbank.org/en/about/legal/terms-of-use-for-datasets

[3] CPRD client application form: https://www.cprd.com/Data-access

Fig. 1. Highlighting important terms to be encoded in a sample of data-governance rules from three sources

be informed of all uses of the `DoB` *column, to prevent harmful disclosure. Apart from that proviso, DP permits the data to be shared with the public, and allows anyone to produce derived work. DP also wants to be credited for producing this dataset by being cited in publications produced by its users. Therefore, they state these two requirements in their data-use policy, and have set up a use-reporting mechanism (*`report.example.ac`*). A data user UA processes the data, and produces two output datasets: DB and DC. DB does not contain* `DoB`*. DC contains only the* `YroB` *(Year of Birth) derived from* `DoB`*. UA wishes to share these two datasets with other researchers.*

Naturally, a few consequences emerge from this example. After obtaining the data D, the data user UA performs data processing independently from DP. As specified in the example, DB does not contain the sensitive information `DoB`, so it would not be bound to the obligation of reporting uses; DC contains derived information `YroB` so it can be considered as still bound to the reporting obligation. Therefore, when UA shares DB and DC to other, appropriate policies for each of them (derived from the original policy) should be attached as well. Similarly, any future users (e.g. UX) using DB is not bound to the reporting obligation too, while users of DC are. If a user UY uses both DB and DC, he/she is still bound to all the original policies (union of the policies of DB and DC), even though UY obtains data from UA instead of DP and might not be aware of the existence of DP. Similar to UA, UX and UY can perform arbitrary processing to the data, creating different consequences for the policies.

From that, we identify the 5 major properties, which are also issues to solve in such contexts, below:

- †1 (Personnel) **Scattering**: data processing is multi-institutional so that data providers and data processors are rarely in the same institutional framework.
- †2 (Rule) **Propagation**: derived data (output data) can be used as input data further, by the same or different people in the current activity or some future activity,
- †3 (Rule) **Diversity**: policies not only impose access control, but also contain general *obligations* that current and future users should fulfil.
- †4 **Dynamic** (rule) **application**: processes change data and therefore can revise / change the policies applied to data, in particular lowering the policy restrictions.
- †5 (Rule) **Combination and separation**: processes can be multi-input-multi-output (MIMO). This may also be checked in two halves:
  - †5.1 (Rule) **Combination**: processes may take multiple inputs with different policies.
  - †5.2 (Rule) **Separation**: processes may produce multiple outputs with different policies.

These identified issues demonstrate the necessity of having automated frameworks to support both the data providers and the data users to deal with rules. Section 2.2 summarizes the different features and focuses on related research taking a similar direction, and concludes that there is a lack of frameworks to solve all 6 identified issues in the federated context.

Therefore, in this paper, we present our work, an intelligent framework called Dr.Aid (Data Rule Aid), which addresses all these aspects and therefore supports data-use rule compliance in a broad range of federated contexts.

The structure for the rest of the document is: the broader background and related work are discussed in Section 2; the introduction to the framework is in Section 3; we present the evaluation in Section 4; after that, in Section 5, we discuss the future work; finally, in Section 6, the conclusion is drawn.

2 BACKGROUND AND RELATED RESEARCH

This section discusses the background that shapes our research goals and reviews related work with similar goals.

2.1 Background

Processing data with the support of computer systems is one of the most common collaboration practices today, particularly for research. This is often denoted as data-intensive research [34], where the role of data sharing is dominant.

The importance of data governance, data ethics and privacy has risen in recent years driven by the widespread application of machine learning [43] and the Internet of Things (IoT) [44, 66], which generate and use massive amounts of data on a daily basis. Connecting this with the so-called "biggest lie on the Internet" [50] (i.e. the fact that most people accept website Terms of Service and Privacy Policies without reading or understanding them) reinforces the same issue whenever people try to enhance their control over data usage, due to the same reason: information overload. Legislative approaches such as the European General Data Protection Regulations (GDPR) bring some consistency and require to give control back to the data subject (normally the user) [1], but they do not eliminate the complexity for people, leaving the issue still open. Therefore, appropriate methods and practical frameworks are needed to facilitate every stakeholder to respect data ethics and governance.

Efforts have been made to address challenges around privacy by algorithmically eliminating the necessity and thus the use of original sensitive data, namely differential privacy [20] (where sensitive-data details are obscured in

Table 1. Summary of framework features regarding our identified issues for realistic contexts where multiple distributed participants progressively import, combine and process data.
✓ means supports; ✗ means does not support; ✓✗ means partially supports; ? means unknown.

Framework	Scattering	Propagation	Diversity	Dynamic application	Combination	Separation
E-P3P[38]	✗	✗	✓	✗	✗	✗
Thoth[29]	✗	✓	✗	✓✗[2]	✓	✗
DAPRECO[26, 57]	?	✗	✓	✗	✗	✗
Smart object[59]	✓	✓	✓	✗	✓	✗
CamFlow[53]	✓	✓	✗	✓	✗	✗
Meta-code[36]	✓✗[3]	✓	✓✗[4]	✓	✗	✗
Dr.Aid (our work)	✓	✓	✓	✓	✓	✓

synthesised derivatives) and federated learning [43] (where sensitive data are restricted to local processing). They provide useful methods in protecting privacy while also keeping high accuracy and personalization. However, issues remain because privacy is not the only element for data governance and ethics. Besides, in many cases, sharing sensitive data is necessary and desired [40], so that decentralized and fine-grained governance is explicitly required.

Some research points out the diversity of people's preferences, and provide automated agents to negotiate with the data accessing body on behalf of the user [25, 37]. This directly addresses the governance and ethics challenges with reduced human effort, particularly in the context of IoT and smart devices with unpredictably many negotiation/authorization requirements. However, they follow a traditional view of data processing where the data is used in one processing step (directly by the organisation to which consent has been granted) or a limited step shared with third-party; data processing has no context and one consent governs forever data usage; derived data products are beyond the scope of control of the consent. Besides, in the general context, data processing can be multi-staged and/or conducted by multiple bodies, and thus goes beyond the limitations of these solutions.

As we describe below (Section 2.2), another research thread focuses on distinct policy requirements and the use of automated frameworks to check and/or ensure compliance. This can reduce human efforts (for data consumers), facilitate the authoring and maintenance of data governance rules (for data providers), and maintain compliance for not only the initial data but also its derivatives. We view this as a necessary direction, such that it may be combined with the automated negotiation agents described above to enable full-fledged practical frameworks maximizing social benefits while also respecting individuals' preferences.

2.2 Related research

In this part, we discuss the related research presenting automated systems to ensure compliance with data governance rules. We discuss them below, and summarize their achievements relative to our five identified issues in Table 1.

One direction of research focuses on ensuring the compliance in a known closed context (e.g. within an institutional boundary). For instance, E-P3P [38] provides a formal model to check compliance before granting access to data. It also enlightened the concept of *sticky policy* [48] (see below). Thoth [29] uses a more flexible logic-based formalisation to encode access control rules as well as automatic declassification conditions, but can not describe *obligations* (required actions as a consequence of using the data) as [38] does. DAPRECO [26, 57] is a legal-modelling approach taking a

[3]Only *declassification* rules which removes compliance checking completely.
[4]Through meta-code, custom arbitrary program code.

similar view, converting legal documents (e.g. EU GDPR, General Data Protection Regulation) to logical expressions and check compliance of some processing. These approaches have different strengths and flexibility, but they hold a narrow view of data processing: data processing seldom affects the applicability of policies. As a result the data-use policy for the input data invariably pertains to all derivatives until the result meets the *declassification* requirement specified by the original policy maker / data governor; the declassification makes the result no longer bound to the original policies, nor to any policies. Sticky policy [48, 54] raised the policy enforcement issue for decentralized contexts, and provided a conceptual framework for maintaining policy compliance in such contexts. [59] (denoted as *smart objects*) provides a model to encode not only the direct data-use policy, but also the mechanism to derive the policies for derived data. Such frameworks are aware of the decentralized context and provide rich controlling power to the data provider. But they require a close collaboration between the data providers and the data users to allow data providers to foresee the processes that the data may go through and encode that in the policies. As a summary, these research constitute useful approaches when the data providers and the data processors are closely collaborated or within the same institutional framework. But for loosely coupled contexts (such as the federated context identified above), it is almost impossible to predetermine the processes the data will go through as methods evolve during the collaboration, and therefore such frameworks could not provide expected support.

Aside from frameworks, there are dedicated policy languages, such as the Open Digital Rights Language (ODRL) [19] and the eXtensible Access Control Markup Language (XACML) [17]. XACML is an XML-based standard used to describe access control; ODRL is a W3C standard based on semantic technologies to describe various aspects of data's terms of use. Regardless of their differences, the primary purpose of these languages is to formally represent data-use policies and check whether a *single* use conforms to them. They do not address rule propagation, data derivation, merging or separation. Thus they possess the same issues as the frameworks discussed above.

A few other research address the propagation and dynamic application issues, explicitly focusing on allowing processes to change the policies associated with derived data. Meta-code [36] and CamFlow [53] utilize concepts from (decentralized) Information Flow Control (IFC) [49] as the foundation to specify the policies and change of policies, and make different extensions. The basic concept is to assign tags to data and specify additional constraints of tags to processes/programs: processes have different input tag compatibility, so only compatible data can be taken as input; processes will produce output, so the tags for output data are specified along with the processes. Meta-code [36] introduced the *meta-code* concept to model the policies that can not be captured by role tags, which are custom program code; CamFlow uses the model of decentralized IFC with two labels (each contains a set of tags), secrecy and integrity, to represent different policy semantics; output policy is specified by manipulating input labels, one label allowed for each process. As a result, Meta-code supports richer types of policies but lacks formality, making static analysis difficult. CamFlow has semantics limited to the two labels, within access control. Both approaches and their developments building on them that we have found do not support MIMO processes.

3 THE DR.AID FRAMEWORK

Our work in this paper, Dr.Aid (Data Rule Aid), is designed for the federated collaboration context. Figure 2 illustrates the general concept of the framework, by connecting data flow with rule flow, addressing the MIMO issue and supporting dynamic rule application.

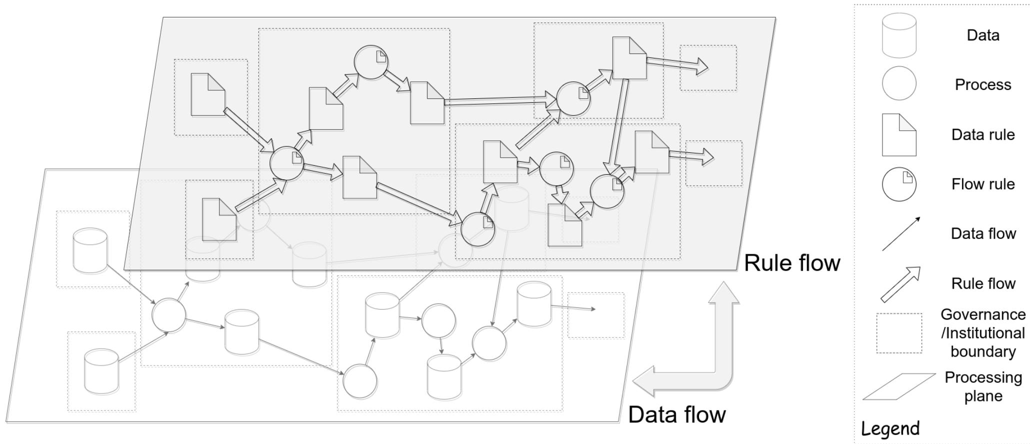

Fig. 2. Conceptual design of the Dr.Aid framework, shifting from the *data flow* at the lower level to the *rule flow* at the upper level

This section presents the design of Dr.Aid framework, including the language and our system architecture. The example described in Section 1 will be revisited along the introduction. More specifically, we assume the user UA uses a process which produces dataset DB from output port[5] $output1$ and dataset DC from output port $output2$.

The language model of Dr.Aid is based on the concept originally proposed by Zhao and Atkinson in [65]. We have revised its design, and addressed additional issues. The major ones are:

- We provided a formal description of the model.
- In addition, we provided a logical interpretation of the model and reasoning mechanism based on a well-studied logic system, namely situation calculus [46, 56]; this also supports whole-graph reasoning (as opposed to process-by-process reasoning).
- We integrated our implementation with a well-known dedicated situation calculus reasoner, Golog [42].
- We provided an abstract intermediate graph model to support compliance checking of provenance from both data-streaming (S-Prov[6] for dispel4py [33]) and file-oriented (CWLProv[7] for CWL[22]) workflow systems.
- We evaluated the model and framework against real-life use cases (in Section 4).

3.1 Design and language

The core concept is to present a formal language containing both the part to model data-use policies (the *data rules*) and the part to model propagation and changes to the data-use policies in processes (the *flow rules*), with a mechanism allowing them to interoperate, and perform reasoning on top of the data-flow graph. This concept is related to decentralized IFC [49], but depicts the general context and supports obligations (actions to be performed after using the data), which makes it different to (D)IFC and other traditional solutions which focus on access controls. Future work can be done to bridge between these two streams of work (see Section 5).

The *data rules*, as a means to model data-governance rules, are associated with data. They contain two main building blocks: *attributes* and *obligations*. An *attribute* describes properties of the data and is represented as a triple (N, T, V)

[5]Processes, the main building blocks of scientific workflows, can take multiple inputs and multiple outputs, each through one of its input ports and output ports.

[6]https://github.com/aspinuso/s-provenance

[7]https://w3id.org/cwl/prov/

of a name N, a type T and a value V. It is the main building block, which is used by obligations, and receives special attention in *flow rules*.

An *obligation* specifies an action (to be performed by the user) such that is triggered under specific conditions, as well as its "dependency" attributes. Formally, an *obligation* is a triple (OD, VB, AC) consisting of an obligation definition OD (the action to perform upon activation), a validity binding set VB (describing additional applicability constraints), and an activation condition AC (the triggering condition). The OD is another tuple (OA, AR) where OA is the obligated action class and AR is a list representing action parameters. In particular, each element of AR and VB refers to an attribute in the data rule (this design is described in *flow rules*). The activation condition AC is a boolean expression, which will be evaluated into true or false with runtime information when checking the activation of obligations (Section 3.2.1). Appendix A summarizes the available *slots* which are the aspects that can be checked (e.g. process type, time of execution, etc).

For instance, the example rule above regarding the reporting of any use of the sensitive "DoB" field to an example URL `report.example.ac` can be modelled as follows:

$$attribute(pf, column\ \text{"}DoB\text{"})$$

$$attribute(ru, url\ \text{"}report.example.ac\text{"})$$

$$obligation(report\ ru, [pf], action = *)$$

Most elements in this formal notation, which we call the "*user notation*" can be directly mapped from the original natural language rules. This is further extended to we call the "*situation calculus representation*" automatically to include additional information used by the situation calculus reasoner during inference (see Section 3.2.3). Figure 3 shows a comparison between different representations. It shall be explained as: the rule segment *column*"*DoB*" is modelled as an attribute whose type is *column*, value is *DoB*, and name is pf (*private field*); the rule segment $url\ report.example.ac$ is modelled as another attribute whose type is *url*, value is `report.example.ac`, and name is *ru* (*report url*); the main content is an obligation declaration with reference to these two attributes, whose obligated action is $report$, action argument is *ru* (referencing the *ru* attribute), validity binding is a list with one element $[pf]$ (referencing the pf attribute), and activation condition is $action = *$ meaning it would activate when the data goes through a process with *any* action type.

The *flow rules*, on the other hand, describe how the data rules would flow through a process, reflecting the underlying data propagation and processing. They involve three types of actions: *propagate*, *edit*, and *delete*. *Propagate* specifies the unedited flow of data rules from input ports to output ports, when no edit or delete is applied. It is a tuple $pr(P_{in}, Ps_{out})$ where P_{in} is the input port to propagate data rules from and Ps_{out} is a set of output ports to propagate data rules to. After specifying propagation, further refinements can be done to the data rules, to reflect the processing and modification of underlying data results in the change of data rules (policies), and specifically the **edit** action and the **delete** action. *Delete* is specified as $delete(P_{in}, P_{out}, N, T, V)$ where P_{in} is an input port, P_{out} is an output port, N is the name of a attribute, T is the type of an attribute and V is the value of an attribute. It acts as a *filter* to match all the data rules (of the process), and remove every matched *attribute*. As a consequence, every *obligation* which refers to these *attributes* (in their action parameters or validity bindings) is removed as well. Similar to delete, *edit* is specified as $edit(P_{in}, P_{out}, N, T, V, T_{new}, V_{new})$ where P_{in}, P_{out}, N, T and V are the same as those in *delete*, T_{new} is the new type of the attribute and V_{new} is the new value of the attribute. The filter is similar to delete, but the matching attributes will

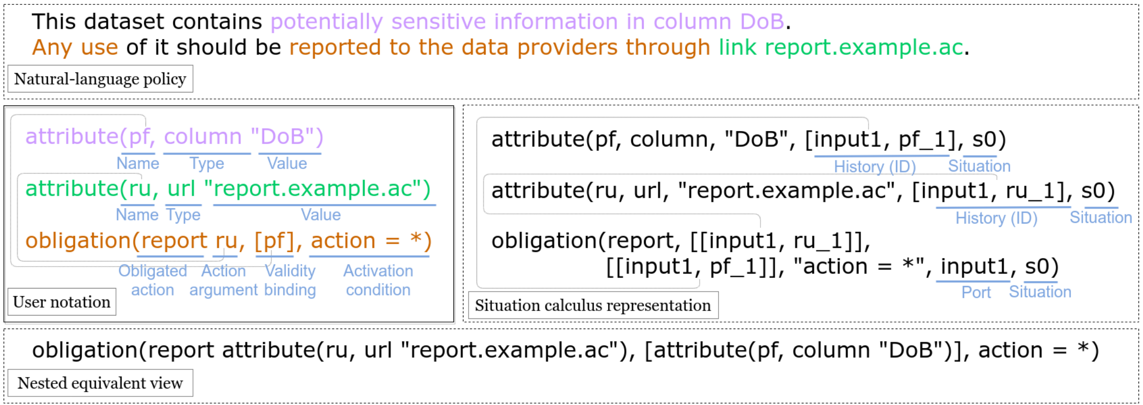

Fig. 3. Encoding (and equivalences) of the example data-governance rule (associated with $input1$)

have their type and value updated to the specified new type T_{new} and new value V_{new}. In addition, each value of the filter (excluding new values) can be specified as a special value $*$, which corresponds to *any possible* value.

For instance, the flow rule for the example process can be specified as (in the user notation):

$$pr(input1, [output1, output2])$$

$$delete(input1, output1, *, column, "DoB")$$

$$edit(input1, output2, *, column, "DoB", column, "YroB")$$

This says the data rules will be propagated from $input1$ to both $output1$ and $output2$, under revision a) to delete attributes from port $input1$ to port $output1$ with *any* ($*$) name, type $column$ and value "DoB", b) to change attributes from port $input1$ to port $output2$ with *any* ($*$) name, type $column$ and value "DoB" to type $column$ and value "$YroB$". By definition of the semantics, the revision a) also deletes any obligations bound to the deleted attributes from $output1$, i.e. the reporting obligation, but it won't affect $output2$.

3.2 Reasoning mechanism

Reasoning is performed by taking the data rules for each input port, executing flow rules, and obtaining the data rules for each output port.

Using the example with the encoding above, the outputs can be automatically calculated to have the following data rules:

Data rules of $output1$ (i.e. of DB).

$$attribute(ru, url\ "report.example.ac")$$

Data rules of $output2$ (i.e. of DC).

$$attribute(pf, column\ "YroB")$$

$$attribute(ru, url\ "report.example.ac")$$

$$obligation(report\ ru, [pf], action = *)$$

Note the dangling attribute *ru* from *output*1 is deliberately kept by the semantics. This design considers the accreditation needs of data providers to leave information in the data rules, and also keeps the language specification simple. While other researchers may prefer to prune the dangling attributes for the sake of simplicity in the data rules, we argue that this is not critical and is merely a design choice.

The reasoning process is intuitive. As demonstrated above, the data rules come in from some input port, which is attached to them during reasoning as necessary information for flow rules; when there are *propagate* rules, the corresponding output ports are associated too, so the *edit* and *delete* can be carried out; after the flow rule processing, the resulting data rules are sent out through the corresponding output ports.

3.2.1 Obligation activation. The procedure above allows us to derive successor data rules. Further reasoning allows checking the activation of obligations. This is done by checking the activation condition of the corresponding data rules at the beginning of each process using contextual information. For the obligations whose activation condition is evaluated to true, their obligation declarations *OD* (including the referenced attributes) will be extracted, and will be put into a separate storage in our implementation. The applied contextual information contains the process information (e.g. process type), the execution information (e.g. the stage during execution) and the provenance information (e.g. the user), as summarized in Appendix A.

3.2.2 Merging and deduplication. Through the flow rules, the rule merging and separation issues is mostly solved – the user is able to explicitly specify how the rules would flow. However, there is still an undiscussed case when different incoming data rules have duplicated entries. Consequently, the output data rules may have duplicated entries propagate (as-is or as the result of editing) if handled naively. Logically, the data rules coming from and going to a port form a set. Therefore, when merging happens, the framework also removes duplicated entries.

3.2.3 Situation calculus formalization. In our work, the language and the reasoning mechanism is provided with a logical background using situation calculus [47], a well-studied logical formalism to characterize dynamic domains, consisting of a decidable extension to first-order logic. Based on these facts, situation calculus is both simple and a good fit for our requirements.

Our method is to align the model components and reasoning with the constructs in situation calculus, which is to model the data rules (plus the associated ports) as *fluents*, the flow rules as *actions*, the different steps of flow rule execution as *situations*, and the reasoning as the *projection task*, i.e. given a target situation (state) S_f, query the fluents that hold in S_f.

The fluent-based situation calculus representation, as shown in Figure 3, contains information about the *history*, i.e. the ports that the information has gone through in each stage, and the current *situation*, i.e. the current state in addition to the parameters of the formal specification discussed earlier.

Due to the particular focus and length consideration of this paper, we do not present the full explanation of this formalization. See Appendix B for the list of relevant axioms (precondition axioms and successor-state axioms).

3.3 System implementation

We built a system implementing the reasoning mechanism above, as well as reading and handling other relevant information. The system is mainly implemented in Python and uses Golog (on SWI-Prolog)[8] as the situation calculus reasoner. Figure 4 gives a high-level view to the architecture of our implementation.

[8]The Golog implementation is obtained from http://www.cs.toronto.edu/cogrobo/main/systems/index.html.

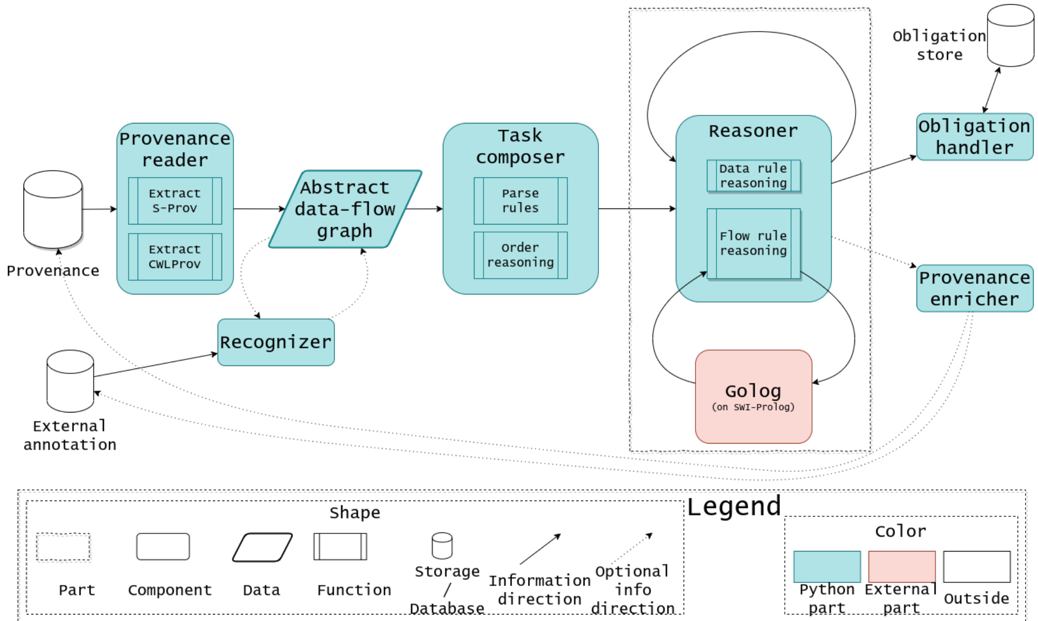

Fig. 4. High-level view of the system architecture

The main goal of the system is to take a data-flow graph whose input data and processes have rules (data rules and flow rules) associated, and to perform reasoning over the data-flow graph to obtain: (1) any activated obligations; (2) data rules associated with output data after the processing. Therefore, in turn, the obtained derived data rules can be used as input data rules for further reasoning.

3.3.1 Input source. The implementation performs retrospective analysis by taking provenance traces as the source of data-flow graphs. The main benefit of provenance is that it allows us to abstract from the implementation details of various (workflow) execution systems, thanks to the standard core ontology, W3C PROV-O [18], and the interoperability provided by the semantic technology.

Our system uses provenance traces produced by scientific workflows [23], which have two major types, file-oriented and data-streaming. For file-oriented workflow systems, each process takes inputs from data files, and produces outputs to data files. The files are either hard-coded in the source code, or passed in as parameters to the processes. On the other hand, processes in data-streaming systems read inputs directly from the outputs of its predecessor processes, without storing to files. The outputs are usually small data units, each representing a meaningful segment of the full output (e.g. a line in a table, a number in a sequence, etc). Such differences give them different capabilities, and also imposes different requirements for the provenance scheme. Because PROV-O is a low-level model, extensions are developed to provide higher-level descriptions for specific needs. In our implementation, we support two provenance schemes for each one of them, namely CWLProv and S-Prov.

In order to support the distinct properties of different schemes, Dr.Aid uses an abstract intermediate representation for the data-flow graph (a visualization example can be found in Figure 5). The main reason we don't use PROV-O directly is because PROV-O is too low-level and causes redundancy in the data production and consumption for data-streaming workflows (S-Prov in our example). In addition, PROV-O is retrospective while our model is not; PROV-O implies the strict existence of intermediate *entity* (e.g. data) between two *activities* (e.g. processes), which can become a limitation in the future to expand the use cases to process graphs without explicit data, e.g. BPMN [51].

3.3.2 Recognizer module. In order to associate rules with the data-flow graphs to cope with the fact that not all data and processes have rules associated with them already, we use the *recognizer* module. Before reasoning, the recognizer checks the data-flow graph, finds matching rules from its database, and injects these extra rules to the data-flow graph. The recognizer also supports identifying processes that need to add additional rules apart from its inputs (e.g. those downloads data internally with no input ports), and inject data rules to such processes. In our implementation, the database is stored as a JSON file.

The database used by the recognizer can also be used to store the reasoning results, i.e. data rules associated with the output data. This is useful for doing experiments, and also useful when the provenance store does not allow to write back (e.g. due to permission issues).

3.3.3 User actions as virtual processes. Inspired by W3C PROV-O, Dr.Aid uniformly treats user actions and computational processes. Therefore, user actions can be injected as *virtual processes*, and the reasoning will go through the same procedure to check activation and/or propagate data rules. In our implementation, this is done by adding extra annotations to the abstract intermediate graph representation to include virtual processes when necessary.

3.3.4 User queries. The implementation has two major user interaction points: (1) Setting the data (provenance and rules) source and execute the reasoning; (2) Checking the activated obligations. Both points are explained above, while the 2nd point is only briefly explained when introducing the activation of obligations. The users are expected to check the activated obligations after the reasoning, and perform actions accordingly. This is enough for experimental purposes as proof-of-concept. In an ideal world, the 1st point can be automatically executed whenever suitable, and the users are expected to check only the 2nd point, through a proper notification mechanism.

4 EVALUATION

In this section, we present the evaluation we performed for Dr.Aid. The evaluation covers:

(1) the ability of our implementation to handle real-world data-flow graphs in collaboration contexts;
(2) the capability of the language for expressing real-world data-governance rules.

Our first two evaluations are based on the use of Dr.Aid in two real-life scientific workflows: cyclone tracking for global-warming impact modelling and Moment Tensor in 3D (MT3D) computing the expected impact of an earthquake. Then we evaluate the capability of the language to specify a selection of diverse real-world published data-governance rules.

4.1 Experimental consistency

Each evaluation has specific properties, but there are commonalities shared between them. The most important one is the procedure to convert from natural-language policies to our formal representation. We have standardised the procedure for this:

(1) Identify and obtain nested rules if any;
(2) Remove unnecessary information from the rules;
(3) Identify *actioning* rules, in particular obligations;

(4) Find the terms in the rules that identify the data or critical properties of data that need to be carried with data, as attributes;
(5) Identify *implied* rules;
(6) Write in the user notation where possible;

The *actioning* rules are the rules that describe an action, which can be an action/behaviour to be complied with when using the data, an action to be performed after using the data, or an action imposed by someone else (usually the data provider) on the user. They are the major contents of rules to be encoded in our model. The *implied* rules are implicit in our model and need not be encoded. An example is *"the user is allowed to redistribute the derived data"*. Implicit behaviours can be explicitly overridden when necessary.

It is worth noting that not every sentence in the natural-language policies can be modelled using our formal language, because those sentences describe contextual information, or because they are beyond the capability of our current model. We discuss such cases as they arise.

Therefore, following this standardized encoding procedure, we measure its effect using the following information:

(1) The total number of sentences in the original natural-language (English) policy;
(2) The total number of rules in the original policy;
(3) The total number of actioning rules;
(4) The total number of implicit rules;
(5) The total number of encoded rules.

4.2 Framework evaluation

The framework evaluation tests the capability of the whole framework with use cases that involve typical collaborative use of data and computational methods for global research addressing environmental hazards [24, 39]. It considers the language encoding, the system implementation, the extracted information, the reasoning result, etc.

As mentioned previously, the selected instances of collaborative behaviour are climate-scientists setting up and running *cyclone tracking* workflows and seismologists setting up and steering workflows to estimate an earthquake's impact in an area they select, either to advise emergency response or to improve regional models for future use (*MT3D*). We use the provenance traces generated by the executions of these workflows, encode the data-use policies of the data selected by users and imported from an open-ended set of providers during these executions, and work with the scientific researchers who authored and executed these workflows to validate our results.

As well as being typical of the collaborative use of data in multi-disciplinary, multi-site loosely coupled federations, we choose these two examples because they contain complex data use patterns, involving multiple processing stages, data separation and data merging. They also illustrate Dr.Aid's applicability for different types of workflow systems, data-streaming with `dispel4py` and task-oriented with CWL. The MT3D workflow consists of multiple sub-workflows set up and individually steered by the seismologists, enabling us to demonstrate that Dr.Aid's compliance checking spans multiple user actions, potentially conducted by different users, in different organisations with arbitrary time separation.

For both use cases, the provenance traces are obtained from SPARQL endpoints served with Apache Jena Fuseki 3.17[9]. We present relevant features of the two applications, and then the results of the two evaluations, which we then discuss.

[9] Apache Jena Fuseki: https://jena.apache.org/documentation/fuseki2/

4.2.1 Cyclone tracking. The cyclone tracking workflow is used to estimate the distribution of tracks of cyclones as a consequence of climate change. It can also track high-pressures and mid-altitude weather systems. Its core component, implemented in Fortran, uses the algorithm and methodology proposed by Sinclair [61]. The workflow is coded in CWL using parallelization (the *scattering* functionality of CWL). Its provenance is delivered compliant with the CWLProv schema. The data used by the original workflow are all obtained from CMIP6[10] whose data-governance rules are presented in [3]. The encoding and discussion are presented in Appendix D, and summarized in Table 2.

4.2.2 MT3D. Moment Tensor in 3D (MT3D) is a seismology use case used to study wave propagation and hazard assessment through characterizing the earthquake properties, including the source parameters and their uncertainties. The Earth is represented in a 3D spectral-element model (SEM) of wave speeds. Unlike cyclone tracking, the MT3D workflow is not a single workflow, but comprises of several sub-workflows which are executed consecutively, with independent provenance traces that need to be correlated. Most of the sub-workflows use `dispel4py`, and provenance traces are in S-Prov schema; while the waveform simulation code, SPECFEM3D [55] is driven using CWL; its provenance is converted to S-Prov by the enactment system. The evaluation performs reasoning on these traces one by one. MT3D has multiple input data for different purposes:

- SEM mesh modelling the Earth's structure;
- The observed earthquake data from seismometers to correlate with model output to estimate errors an iteratively improve the source model;
- Initial parameters identifying the earthquake source.

They can come from different sources, e.g. EIDA[11], INGV[12], Global CMT Catalogue[13], etc. In our experiment, the mesh and wavespeed profiles for the SEM modelling are obtained from personal communications, the observed earthquake data are from EIDA, and parameters of earthquake source are from INGV. The policy for the personal communication, as we obtained from the workflow's author, was a requirement to properly acknowledging the data provider. The properties and encoding of the publicly available policies are summarized in Table 2 in Section 4.3; the policy for the personal communication is encoded but not included in that table. The encodings and their justifications are presented in Appendix E. As a summary, the model was successful in encoding all of the actioning rules.

4.2.3 Result and discussion. For the cyclone tracking workflow, Figure 5 shows the identified data-flow graph and activated obligations. The top diagram is the visualization of the original data-flow graph (in our intermediate representation) extracted from the provenance; it receives some extra annotations (e.g. the magnified part) to clarify important aspects; the printed dictionary at the bottom is the identified activated obligations; the table to the right is the information stored in the obligation database, corresponding to the dictionary result at the bottom. Readers may identify some repetitions which are expected because of the semantics: the data are used in parallel, and therefore each trace creates one activation following the definition in the data rules (more precisely, the activation condition stage = import). It is an open question whether to keep them, deduplicate them, or to provide another mechanism for specifying them in the rules, which is beyond this paper. As a quick solution, in a deployed system, a user-interface may present the logically distinct obligations. In addition to the identified activated obligations, the reasoning result also contains the derived data rules for each output data. That is shown in Figure 6 in Appendix C.1.

[10]CMIP6 website:https://pcmdi.llnl.gov/CMIP6/
[11]EIDA: http://www.orfeus-eu.org/data/eida/
[12]INGV: http://www.ingv.it/
[13]Global CMT Catalogue: https://www.globalcmt.org/CMTsearch.html

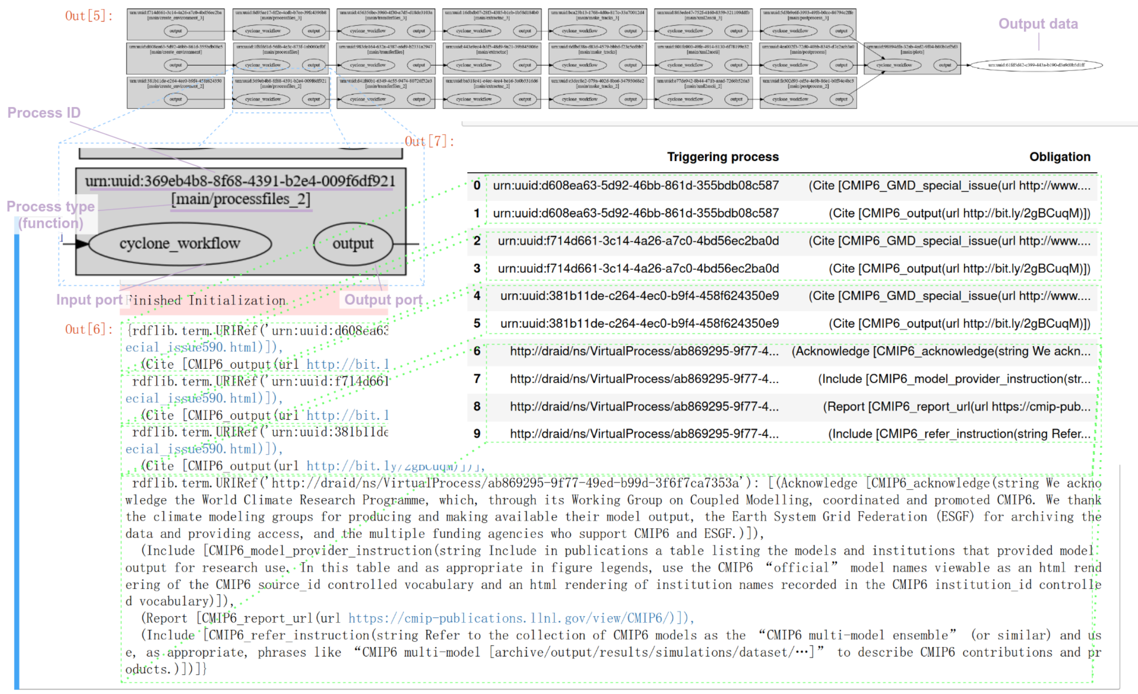

Fig. 5. Visualization of and identified obligations from reasoning about the data-flow graph of the cyclone-tracking workflow.

The reasoning results for MT3D are also as expected. The difference is that it has multiple sub-workflows, and therefore the provenance traces are also separated. As a result, the reasoning needs to be run for each of the traces. The reasoning results from the previous traces are then fed into the subsequent traces (where relevant). The prevailing data rules associated with derived data are retained until the end. Because some data rules use the trigger of process = publish, the required obligations are triggered at the "publish" virtual process. In addition, as shown in Figure 9, the left-most processes explicitly uses flow rules to regulate the flow of rules from the expected input ports to the corresponding output ports. The figures showing the reasoning results are shown in Appendix C.2.

As a summary, our framework correctly extracted the necessary information from the provenance traces and correctly carried out the reasoning of the flow rules and data rules, including executing flow rules, associating the expected data rules with data products and triggering expected obligations at the correct time. We conclude that our implementation addresses the rule-handling issues we have identified effectively and has the potential to do this in a wide range of deployments.

4.3 Encoding real-world public data-use policies

To evaluate our model's wider applicability, we examine it using published data-use policies. The main focus is the capability of our model, that is to what extent can our model represent the rules of those policies. We evaluate that by encoding them in our formal representation, and compare our formalization with the original policy.

Table 2. Result summary for 15 published data-use policies showing the coverage of the formalization.

Policy source	# sentences	# rules	# actioning	# implied	# encoded	actioning coverage	total coverage
CMIP6[3]	35	9	8	1	7	100%	89%
EIDA[8]	20	5	3	0	3	100%	60%
INGV[7]	2	2	2	0	2	100%	100%
CC-BY[5]	12	6	5	2	3	100%	83%
CMT Catalogue[9]	15	4	4	0	4	100%	100%
CORDEX[4]	22	9	6	0	5	83%	55%
ISMD[12]	2	1	1	0	1	100%	100%
RCMT[10]	14	3	3	2	1	100%	100%
MIMIC[13]	17	4	4	0	4	100%	100%
CPRD[6]	21	7	6	0	2	33%	29%
PIMA[15]	2	1	1	0	1	100%	100%
ISC[2]	21	7	7	0	7	100%	100%
IRIS[11]	28	10	10	0	10	100%	100%
OGL[14]	30	7	4	3	1	100%	57%
World Bank[16]	40	12	7	2	3	71%	42%
Total	281	87	71	10	54	90%	74%

4.3.1 Evaluation design. We first identify and collect published data-use policies from a range of data providers and archival services typical of the resources used by practitioners working on data-driven research. These *policies* are publicly available, and the data they govern are often also publicly available, though not always (e.g. MIMIC [13]). Then, we follow our standard procedure to convert from the natural-language policies to our formal representation. We record the results, and provide different metrics supporting comparison; Table 2 summarizes these. We then present our interpretation of this evaluation.

Policy origin. The policies were collected from publicly available sources. These were found by asking the research scientists what dataset they would use and tracking back to find the relevant data-use policies. We also navigated to the related datasets that these services referenced (e.g. by following the link on their website). Another source was searching for datasets and policies on the Internet. (Contrary to our intuition, the latter method did not produce many useful results.) It is also worth noting that the collected target policies are the data-use policies for the *data users* to comply with. Such policies may have a backing legal formality, but that formality is not our target.

Figures and Metrics. Based on the information, the two main indices we evaluate on are:

(1) actioning rule coverage: the proportion of actioning rules in the policy that are encoded;
(2) total rule coverage: the proportion for all rules, not limited to actioning rules, of the policy that are encoded.

And they are defined as:

$$\text{actioning rule coverage:} \quad C_{act} = \frac{N_{enc} + N_{imp}}{N_{act}}$$

$$\text{total rule coverage:} \quad C_{tot} = \frac{N_{enc} + N_{imp}}{N_{rule}}$$

where N_{enc} is the number of rules encoded, N_{imp} is the number of rules implied, N_{act} is the number of actioning rules, and N_{tot} is the total number of rules.

4.3.2 Result and discussion. The results are summarized in Table 2, and the encodings are available in Appendix F (and Appendix D for the ones from cyclone tracking, E for the ones from MT3D). As can be seen, our model is able to represent a high amount of actioning rules (with $\sum C_{act} \approx 90\%$). This demonstrates that our model is in a valid direction to characterize real-world data-use policies. It does not reach 100% because of the limitation of the current semantics – only *obligations* are supported. It has a lower rate in representing non-actioning rules (with $\sum C_{tot} \approx 74\%$), because of the emphasis of the framework. The primary design goal was to help users comply with policies and share derived data respecting those policies, so the contextual and disclaimer rules are not included in the current model. This can be solved by extending the semantics to include such rules; planned in our future work. Digging into the details, we have some additional findings discussed below.

Acknowledgement. All policies present the need for proper acknowledgement of the data author or provider (and/or dataset, data service) in subsequent publications, and some of them have multiple acknowledgement requirements. Our model is able to encode such requirements easily as obligations. The activation conditions are well represented in our model as it models user actions as virtual processes and treats them uniformly with computational processes.

Nested policies. Many policies have nested policies which refer to another policy in addition to the rules stated directly. This enhances the usefulness of the automated framework to facilitate compliance, because such nested policies can be automatically included.

Types of rules successfully modelled. In addition to that, our model can represent actioning requirements such as limiting the purpose of use, user of use, and actions about derived data. They cover the majority of the actioning policies we reviewed. In addition, with flow rules, our model is able to specify contextual constraints (which are to be removed with flow rules using $delete()$ in appropriate processes) and changing contextual information for obligations (to be changed by flow rules using $edit()$).

Use of derived data. Most policies don't explicitly specify the extent to which they apply to derived data. Only CC-BY, OGL and World Bank (and the policies include them as nested policies) explicitly specify that they allow the user to redistribute derived data. But considering the context, all data providers do not object to the users to redistribute the derived data, except MIMIC and CPRD which are both medical data. In fact, the data providers for MIMIC and CPRD also do not object to users publishing results using their data, because they have the acknowledgement requirements in their policies. Because the main reason of MIMIC not being publicly available is *"...database, although de-identified, still contains detailed information regarding the clinical care of patients, so must be treated with appropriate care and respect".* We suspect that the reason for not directly allowing derived data to be shared is due to concerns over revealing sensitive information. This links to the example use case we illustrated, and our framework provides a promising direction.

Data merging and CPRD. Our model accommodates data merging as a *by-default* permitted action. This is invariably true for data policies because of their generic role supporting any research or enquiry. It is not true for the CPRD policy, which targets more specific uses of its data subset. Two more rules can be partially modelled if we use *ad hoc* methods, raising the coverage to 66% and 57% for that data provider.

Reflective actions. Another drawback is that our model is unable to represent the commitments that users are *required to make* (and initiated by the data providers, in contract to obligations initiated by the data users), such as 'to provide information on how they used data when required' (e.g. CORDEX, CPRD). This is a potentially useful and straightforward extension in the future, by including action initiator in the obligation.

Compression. As can be observed from the table, the number of total sentences in the original natural-language policies is much larger than the number of rules $\frac{87}{281} \approx 31\%$ (or $\frac{71}{281} \approx 25\%$ if considering only the actioning rules). This shows that information density is not very high in the natural-language policies, due to various reasons, e.g. the necessity to clarity terms, the inclusion of contextual information, duplicated statements, etc. In principle, some of these information can be defined once and shared across policies, but the practice does not follow that. Therefore, even if we just consider the compression of policies, it is already a sensible approach to model the rules using a formal language.

5 FUTURE WORK

We consider the Dr.Aid framework a big step forward, but many issues still need to be addressed, including those exposed by our research. In this section, we present our future work plans.

User study. We have performed a pilot user study with four research scientists in diverse domains. It broadened our view of the target context, identified real-life but often unencoded data-use policies, and strengthened the evidence motivating and shaping our framework. We are preparing a larger study investigating the usability, utility and understandability of the system and language for more of the roles involved in data-intensive research collaborations. The intended subjects will be scientific workers whose work involves data processing, data handling and curation, method development and evidence production. We are obtaining ethical approval and will then recruit participants and conduct the study.

Link with (D)IFC. The language model is related to (D)IFC (decentralized Information Flow Control) [49], but we take a different direction – DIFC binds semantic meanings to the tags directly, while our language separates the controlling element (the *attributes*) and the semantic element (the *obligations*). This results in a more flexible and extensible language. Using this extensibility, we plan to establish a more formal link with (D)IFC. We also intend to draw on, and if possible interact with, other research that is clarifying concepts, developing ontologies and investigating languages that describe data rules and their application e.g. [28].

Extend language. The language currently only supports obligations as the actioning construct. This can be extended in several directions. For example, we can support *restrictions*, preventing the use of data when a condition is not satisfied, which is a useful semantics for extending the coverage of data-governance rules. This is similar to the *pre-obligation* in some other research [30, 52] (they would refer to the obligations in our research as *post-obligations* or *ongoing-obligations*). Extension can also be made on the activation conditions to support complex conditions. More expressive logic constructs may be needed, which may require us to adopt additional logical foundations.

Support logic deduction and optimization. We have used situation calculus as the formal foundation for our language model, but did not investigate its potential extensions. Further optimization can be done (e.g. [31, 62]), such that it would be possible to recursively deduce the "flow rules" of the whole workflow graph from the flow rules of individual processes. That would allow the whole workflow to be treated as a single process when users are not concerned about associating rules with intermediate results; this also enables automatic deduction of flow rules from code level.

6 CONCLUSION

In this paper, we identified and addressed an important and urgent need to supply automation to help workers comply with data-use rules, particularly when they collaborate over long periods and in geographically distributed loosely coupled federations. This computer-supported approach will also help practitioners in simpler contexts. We have shown that the data-rules are widespread and have observed that there is a severe shortage of tools to help users find the relevant rules and comply with them at the critical moments. We clarify the requirement by identifying five important objectives for enabling data-rule compliance in federated contexts. Drawing on relevant contemporary research we opened up a general approach by prototyping a framework, Dr.Aid, which successfully addresses all five objectives. We demonstrated this success using two real-world scientific workflows from meteorology and computational seismology. We also assessed our coverage by encoding the rules published by 15 data repositories. This revealed some limitations and motivated our future work.

We believe this is a major step towards a future where all those involved in data use are supported by a framework inspired by Dr.Aid, covering virtually all data-rules and capturing information from the majority of tools and processes so that the framework can be widely deployed. Humans still take responsibility for formulating rules, but with the improved precision and compliance, rules will become more subtle. Data analysts will be reminded of their obligations as they produce results and as data is passed between them. They retain their autonomy, when they want to they selectively review the unfilled obligations the system has collected for them, drill into details and decide which ones they should deal with. Workflow and software developers understand how their products propagate rules, and will specify when rules can be relaxed as a result of processing or when new rules should be added. Administrators and managers can review obligations. Governance can focus on where rules need revision. This depends on two crucial advances: (1) a formal notation for rules that has a form that users understand and can use, and a form that reasoners can understand and use; and (2) a reasoning system that is coupled to the data handling and processing systems in use that delivers relevant information tuned to each role in the data-sharing community.

REFERENCES

[1] [n.d.]. Chapter 3 – Rights of the data subject. https://gdpr-info.eu/chapter-3/

[2] [n.d.]. Citing the ISC. http://www.isc.ac.uk/citations/

[3] [n.d.]. CMIP6 Terms of Use. https://pcmdi.llnl.gov/CMIP6/TermsOfUse/TermsOfUse6-1.html

[4] [n.d.]. CORDEX Data access. http://www.cordex.org/data-access/

[5] [n.d.]. Creative Commons — Attribution 4.0 International — CC BY 4.0. https://creativecommons.org/licenses/by/4.0/

[6] [n.d.]. Data access | CPRD. https://www.cprd.com/Data-access

[7] [n.d.]. Earthquake List with real-time updates » INGV Osservatorio Nazionale Terremoti. http://cnt.rm.ingv.it/en

[8] [n.d.]. EIDA Data Policy. http://www.orfeus-eu.org/data/eida/acknowledgements/

[9] [n.d.]. Global Centroid Moment Tensor Project Citation Information. https://www.globalcmt.org/CMTcite.html

[10] [n.d.]. INGV - RCMT. http://rcmt2.bo.ingv.it/

[11] [n.d.]. IRIS Citations | IRIS. https://www.iris.edu/hq/iris_citations

[12] [n.d.]. ISMD - Citation. http://ismd.mi.ingv.it/citation.php

[13] [n.d.]. MIMIC Dataset Acknowledgements. https://mimic.physionet.org/about/acknowledgments/

[14] [n.d.]. Open Government Licence. http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

[15] [n.d.]. Pima Indians Diabetes Database | Kaggle. https://www.kaggle.com/uciml/pima-indians-diabetes-database

[16] [n.d.]. Terms of Use for Datasets. https://www.worldbank.org/en/about/legal/terms-of-use-for-datasets

[17] 2013. eXtensible Access Control Markup Language (XACML) Version 3.0. https://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html

[18] 2013. PROV-O: The PROV Ontology. https://www.w3.org/TR/2013/REC-prov-o-20130430/

[19] 2018. ODRL Information Model 2.2. https://www.w3.org/TR/odrl-model/

[20] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*. Association for Computing

Machinery, New York, NY, USA, 308–318. https://doi.org/10.1145/2976749.2978318

[21] Imtiaz Ahmad, Rosta Farzan, Apu Kapadia, and Adam J. Lee. 2020. Tangible Privacy: Towards User-Centric Sensor Designs for Bystander Privacy. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 116:1–116:28. https://doi.org/10.1145/3415187

[22] Peter Amstutz, Michael R. Crusoe, Nebojša Tijanić, Brad Chapman, John Chilton, Michael Heuer, Andrey Kartashov, Dan Leehr, Hervé Ménager, Maya Nedeljkovich, Matt Scales, Stian Soiland-Reyes, and Luka Stojanovic. 2016. Common Workflow Language, v1.0. (July 2016). https://doi.org/10.6084/m9.figshare.3115156.v2 Publisher: figshare.

[23] Malcolm Atkinson, Sandra Gesing, Johan Montagnat, and Ian Taylor. 2017. Scientific workflows: Past, present and future. *Future Generation Computer Systems* 75 (Oct. 2017), 216 – 227. https://doi.org/10.1016/j.future.2017.05.041

[24] Malcolm P. Atkinson, Rosa Filgueira, Iraklis A. Klampanos, Antonis Koukourikos, Amrey Krause, Federica Magnoni, Christian Pagé, Andreas Rietbrock, and Alessandro Spinuso. 2019. Comprehensible Control for Researchers and Developers Facing Data Challenges. In *15th International Conference on eScience, eScience 2019, San Diego, CA, USA, September 24-27, 2019*. IEEE, 311–320. https://doi.org/10.1109/eScience.2019.00042

[25] Tim Baarslag, Alper T. Alan, Richard Gomer, Muddasser Alam, Charith Perera, Enrico H. Gerding, and m.c. schraefel. 2017. An Automated Negotiation Agent for Permission Management. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS '17)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 380–390. http://dl.acm.org/citation.cfm?id=3091125.3091184

[26] Cesare Bartolini, Gabriele Lenzini, and Livio Robaldo. 2019. The DAta Protection REgulation COmpliance Model. *IEEE Security & Privacy* 17, 6 (Nov. 2019), 37–45. https://doi.org/10.1109/MSEC.2019.2937756

[27] Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Danius Michaelides, Stuart Owen, David Newman, Shoaib Sufi, and Carole Goble. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems* 29, 2 (Feb. 2013), 599–611. https://doi.org/10.1016/j.future.2011.08.004

[28] Daniel J. Dougherty, Kathi Fisler, and Shriram Krishnamurthi. 2007. Obligations and Their Interaction with Programs. In *Computer Security – ESORICS 2007 (Lecture Notes in Computer Science)*, Joachim Biskup and Javier López (Eds.). Springer, Berlin, Heidelberg, 375–389. https://doi.org/10.1007/978-3-540-74835-9_25

[29] Eslam Elnikety, Aastha Mehta, Anjo Vahldiek-Oberwagner, Deepak Garg, and Peter Druschel. 2016. Thoth: Comprehensive Policy Compliance in Data Retrieval Systems. In *Proceedings of the 25th USENIX Conference on Security Symposium (SEC'16)*. USENIX Association, Berkeley, CA, USA, 637–654. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/elnikety

[30] Yehia Elrakaiby, Frédéric Cuppens, and Nora Cuppens-Boulahia. 2012. Formal enforcement and management of obligation policies. *Data & Knowledge Engineering* 71, 1 (Jan. 2012), 127–147. https://doi.org/10.1016/j.datak.2011.09.001

[31] Christopher James Ewin. 2018. Optimizing projection in the situation calculus. (2018). http://minerva-access.unimelb.edu.au/handle/11343/219204 Accepted: 2018-12-04T23:16:35Z.

[32] Sebastian S. Feger, Paweł W. Wozniak, Lars Lischke, and Albrecht Schmidt. 2020. 'Yes, I comply!': Motivations and Practices around Research Data Management and Reuse across Scientific Fields. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 141:1–141:26. https://doi.org/10.1145/3415212

[33] Rosa Filguiera, Iraklis Klampanos, Amrey Krause, Mario David, Alexander Moreno, and Malcolm Atkinson. 2014. Dispel4Py: A Python Framework for Data-intensive Scientific Computing. In *Proceedings of the 2014 International Workshop on Data Intensive Scalable Computing Systems (DISCS '14)*. IEEE Press, Piscataway, NJ, USA, 9–16. https://doi.org/10.1109/DISCS.2014.12

[34] Anthony J. G. Hey (Ed.). 2009. *The fourth paradigm: data-intensive scientific discovery.* Microsoft Research, Redmond, Washington.

[35] L. Hutton and T. Henderson. 2018. Toward Reproducibility in Online Social Network Research. *IEEE Transactions on Emerging Topics in Computing* 6, 1 (Jan. 2018), 156–167. https://doi.org/10.1109/TETC.2015.2458574

[36] Håvard D. Johansen, Eleanor Birrell, Robbert van Renesse, Fred B. Schneider, Magnus Stenhaug, and Dag Johansen. 2015. Enforcing Privacy Policies with Meta-Code. In *Proceedings of the 6th Asia-Pacific Workshop on Systems (APSys '15)*. ACM Press, Tokyo, Japan, 1–7. https://doi.org/10.1145/2797022.2797040

[37] Catholijn M. Jonker, Valentin Robu, and Jan Treur. 2007. An Agent Architecture for Multi-attribute Negotiation Using Incomplete Preference Information. *Autonomous Agents and Multi-Agent Systems* 15, 2 (Oct. 2007), 221–252. https://doi.org/10.1007/s10458-006-9009-y

[38] Günter Karjoth, Matthias Schunter, and Michael Waidner. 2002. Platform for Enterprise Privacy Practices: Privacy-Enabled Management of Customer Data. In *Privacy Enhancing Technologies (Lecture Notes in Computer Science)*. Springer, Berlin, Heidelberg, 69–84. https://doi.org/10.1007/3-540-36467-6_6

[39] Iraklis A. Klampanos, Chrysoula Themeli, Alessandro Spinuso, Rosa Filgueira, Malcolm Atkinson, André Gemünd, and Vangelis Karkaletsis. 2020. DARE Platform a Developer-Friendly and Self-Optimising Workflows-as-a-Service Framework for e-Science on the Cloud. *Journal of Open Source Software* 5, 54 (2020), 2664. https://doi.org/10.21105/joss.02664

[40] Nadin Kökciyan and Pınar Yolum. 2017. Context-Based Reasoning on Privacy in Internet of Things. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*. 4738–4744. https://www.ijcai.org/proceedings/2017/660

[41] Janne Lahtiranta, Sami Hyrynsalmi, and Jani Koskinen. 2017. The False Prometheus: Customer Choice, Smart Devices, and Trust. *SIGCAS Comput. Soc.* 47, 3 (Sept. 2017), 86–97. https://doi.org/10.1145/3144592.3144601

[42] Hector J. Levesque, Raymond Reiter, Yves Lespérance, Fangzhen Lin, and Richard B. Scherl. 1997. GOLOG: A logic programming language for dynamic domains. *The Journal of Logic Programming* 31, 1 (1997), 59 – 83. https://doi.org/10.1016/S0743-1066(96)00121-5

[43] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. 2020. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine* 37, 3 (May 2020), 50–60. https://doi.org/10.1109/MSP.2020.2975749 Conference Name: IEEE Signal Processing Magazine.

[44] Y. Li, W. Dai, Z. Ming, and M. Qiu. 2016. Privacy Protection for Preventing Data Over-Collection in Smart City. *IEEE Trans. Comput.* 65, 5 (May 2016), 1339–1350. https://doi.org/10.1109/TC.2015.2470247

[45] Bradley Malin, David Karp, and Richard H. Scheuermann. 2010. Technical and Policy Approaches to Balancing Patient Privacy and Data Sharing in Clinical and Translational Research. *Journal of Investigative Medicine* 58, 1 (Jan. 2010), 11–18. https://doi.org/10.2310/JIM.0b013e3181c9b2ea

[46] John McCarthy. 1963. *Situations, Actions, and Causal Laws.* Technical Report. STANFORD UNIV CA DEPT OF COMPUTER SCIENCE. https://apps.dtic.mil/sti/citations/AD0785031 Section: Technical Reports.

[47] John McCarthy. 1969. *Some philosophical problems from the standpoint of artificial intelligence.* University, Edinburgh.

[48] M. C. Mont, S. Pearson, and P. Bramhall. 2003. Towards accountable management of identity and privacy: sticky policies and enforceable tracing services. In *14th International Workshop on Database and Expert Systems Applications, 2003. Proceedings.* 377–382. https://doi.org/10.1109/DEXA.2003.1232051

[49] Andrew C. Myers and Barbara Liskov. 1997. A Decentralized Model for Information Flow Control. In *Proceedings of the Sixteenth ACM Symposium on Operating Systems Principles (SOSP '97).* ACM, New York, NY, USA, 129–142. https://doi.org/10.1145/268998.266669

[50] Jonathan A. Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (Jan. 2020), 128–147. https://doi.org/10.1080/1369118X.2018.1486870

[51] Object Management Group. 2013. Business Process Model and Notation (BPMN), Version 2.0.2. https://www.omg.org/spec/BPMN/.

[52] Jaehong Park and Ravi Sandhu. 2004. The UCON $_{ABC}$ usage control model. *ACM Transactions on Information and System Security* 7, 1 (Feb. 2004), 128–174. https://doi.org/10.1145/984334.984339

[53] Thomas F. J.-M. Pasquier, Jatinder Singh, David Eyers, and Jean Bacon. 2017. CamFlow: Managed Data-sharing for Cloud Services. *IEEE Transactions on Cloud Computing* 5, 3 (July 2017), 472–484. https://doi.org/10.1109/TCC.2015.2489211 arXiv: 1506.04391.

[54] S. Pearson and M. Casassa-Mont. 2011. Sticky Policies: An Approach for Managing Privacy across Multiple Parties. *Computer* 44, 9 (Sept. 2011), 60–68. https://doi.org/10.1109/MC.2011.225

[55] D. Peter, D. Komatitsch, Y. Luo, R. Martin, N. Le Goff, E. Casarotti, P. Le Loher, F. Magnoni, Q. Liu, C. Blitz, T. Nissen-Meyer, P. Basini, and J. Tromp. 2011. Forward and adjoint simulations of seismic wave propagation on fully unstructured hexahedral meshes. 186 (2011), 721–739.

[56] Raymond Reiter. 1991. The frame problem in situation the calculus: a simple solution (sometimes) and a completeness result for goal regression. In *Artificial intelligence and mathematical theory of computation: papers in honor of John McCarthy.* Academic Press Professional, Inc., USA, 359–380.

[57] Livio Robaldo and Xin Sun. 2017. Reified Input/Output logic: Combining Input/Output logic and Reification to represent norms coming from existing legislation. *Journal of Logic and Computation* 27, 8 (Dec. 2017), 2471–2503. https://doi.org/10.1093/logcom/exx009

[58] David Robertson, Fausto Giunchiglia, Stephen Pavis, Ettore Turra, Gabor Bella, Elizabeth Elliot, Andrew Morris, Malcolm Atkinson, Gordon McAllister, Areti Manataki, Petros Papapanagiotou, and Mark Parsons. 2016. Healthcare data safe havens: towards a logical architecture and experiment automation. *The Journal of Engineering* 2016, 11 (Oct. 2016), 431–440. https://doi.org/10.1049/joe.2016.0170

[59] Gokhan Sagirlar, Barbara Carminati, and Elena Ferrari. 2018. Decentralizing privacy enforcement for Internet of Things smart objects. *Computer Networks* 143 (Oct. 2018), 112–125. https://doi.org/10.1016/j.comnet.2018.07.019

[60] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini. 2015. Security, privacy and trust in Internet of Things: The road ahead. *Computer Networks* 76 (Jan. 2015), 146–164. https://doi.org/10.1016/j.comnet.2014.11.008

[61] Mark R. Sinclair. 2004. Extratropical Transition of Southwest Pacific Tropical Cyclones. Part II: Midlatitude Circulation Characteristics. *Monthly Weather Review* 132, 9 (Sept. 2004), 2145–2168. https://doi.org/10.1175/1520-0493(2004)132<2145:ETOSPT>2.0.CO;2 Publisher: American Meteorological Society Section: Monthly Weather Review.

[62] Michael Thielscher. 1999. From situation calculus to fluent calculus: State update axioms as a solution to the inferential frame problem. *Artificial Intelligence* 111, 1-2 (July 1999), 277–299. https://doi.org/10.1016/S0004-3702(99)00033-8

[63] Tim Berners-Lee. 2009. Linked Data - Design Issues. https://www.w3.org/DesignIssues/LinkedData.html

[64] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (March 2016), 160018. https://doi.org/10.1038/sdata.2016.18

[65] Rui Zhao and Malcolm Atkinson. 2019. Towards a Computer-Interpretable Actionable Formal Model to Encode Data Governance Rules. In *Proceedings of 2019 15th International Conference on eScience (eScience).* 594–603. https://doi.org/10.1109/eScience.2019.00082

[66] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. 2018. User Perceptions of Smart Home IoT Privacy. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–20. https://doi.org/10.1145/3274469 arXiv: 1802.08182.

A ACTIVATION CONDITION SLOTS

The summary of slots in activation conditions is presented in Table 3. The value can be any value literal or a special constant $*$ representing *any* value.

Table 3. Slots of activation conditions

Slot	From	Meaning
action	Provenance	The process *type*
stage	Framework	The processing stage that this rule is involved
purpose	User specification	The purpose of this workflow execution
user	Provenance	The user identifier, retrieved from the provenance
startTime	Provenance	The date and time of execution
processId	Provenance	The ID of the process

The available values for the "stage" slot are "start-of-workflow" (start of the workflow), "end-of-workflow" (when the workflow finishes)) and "import" (when the rule is imported to the execution for the first time).

B AXIOMS FOR THE SITUATION CALCULUS FORMALIZATION

All the *fluents* are listed here:

$$\begin{aligned} &Attr(n, t, v, h, s) \\ &PropAttr(n, t, v, h, s) \\ &Obligation(ob, h, cond, p_{in}, s) \\ &PropObligation(ob, h, cond, p_{in}, p_{out}, s) \end{aligned} \tag{1}$$

All the *actions* are:

$$\begin{aligned} &pr(p_{in}, ps_{out}) \\ &edit(\underline{n}, \underline{t}, \underline{v}, t_{new}, v_{new}, \underline{p_{in}}, \underline{p_{out}}) \\ &delete(\underline{n}, \underline{t}, \underline{v}, \underline{p_{in}}, \underline{p_{out}}) \end{aligned} \tag{2}$$

where the underscore marks that this argument may be $*$ which denotes *arbitrary*. We require that the original rule can not contain $*$ as its value for these arguments.

The precondition axioms are simply $\top$ (true), because we expect the action be still perform-able but does nothing when the expected conditions do not hold.

$$\begin{aligned} &Poss(pr(p_{in}, ps_{out}), s) \Leftrightarrow \top \\ &Poss(edit(\underline{n}, \underline{t}, \underline{v}, t_{new}, v_{new}, \underline{p_{in}}, \underline{p_{out}}), s) \Leftrightarrow \top \\ &Poss(delete(\underline{n}, \underline{t}, \underline{v}, \underline{p_{in}}, \underline{p_{out}}), s) \Leftrightarrow \top \end{aligned} \tag{3}$$

The successor-state axioms are:

$$
\begin{aligned}
&PropAttr(n,t,v,h=[h_0|[p_{in},p_{out}]],do(a,s) \Leftrightarrow \\
&\quad PropAttr(n,t,v,h,s) \\
&\qquad \wedge \neg(a = delete(\underline{n},\underline{t},\underline{v},\underline{p_{in}},\underline{p_{out}}) \\
&\qquad\quad \vee \exists v_2 \neq v.a = edit(\underline{n},\underline{t},\underline{v},t_2,v_2,\underline{p_{in}},\underline{p_{out}}) \\
&\qquad\quad \vee a = end(p_{out})) \\
&\quad \vee PropAttr(n,t_{old},v_{old},h,s) \wedge (a = edit(\underline{n},\underline{t_{old}},\underline{v_{old}},t,v,\underline{p_{in}},\underline{p_{out}})) \\
&\quad \vee Attr(n,t,v,h_1=[h_0|[p_{in}]],s) \wedge (a = pr(p_{in},ps_{out})) \wedge p_{out} \in ps_{out}
\end{aligned} \tag{4}
$$

$$
\begin{aligned}
&PropObligation(ob,h=[h_0|[p_{in},p_{out}]],cond,p_{in},p_{out},do(a,s)) \Leftrightarrow \\
&\quad \neg(\exists n,t,v,p_{in},p_{out}.\{PropAttr(n,t,v,h,s) \wedge a = delete(\underline{n},\underline{t},\underline{v},\underline{p_{in}},\underline{p_{out}})\} \\
&\quad \vee a = end(p_{out})) \\
&\quad \vee Obligation(ob,h_1=[h_0|[p_{in}]],cond,p_{in},s) \wedge a = pr(p_{in},ps_{out}) \wedge p_{out} \in ps_{out}
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
&Attr(n,t,v,h=[\_|[p]],do(a,s)) \Leftrightarrow \\
&\quad Attr(n,t,v,h,s) \wedge \neg\exists psa = pr(p,ps) \\
&\quad \vee PropAttr(n,t,v,h,s) \wedge a = end(p)
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
&Obligation(ob,h,cond,p,do(a,s)) \Leftrightarrow \\
&\quad Obligation(ob,h,cond,p,s) \wedge \neg\exists ps.a = pr(p,ps) \\
&\quad \vee PropObligation(ob,h,cond,p,s) \wedge a = end(p)
\end{aligned} \tag{7}
$$

The arguments correspond to those explained in Section 3.1, so we omit the explanation for simplicity. In these axioms, we use a notation similar to Prolog's notation of lists when retrieving elements in histories, but we do this in the reversed order to indicate that they are appended, conceptually. Similarly, the = in the head/consequence (e.g. $h = [\_|[p]]$) means expansion (to be used later in the body), rather than assignment.

C RESULTS OF FRAMEWORK EVALUATION

C.1 Cyclone tracking

Here we have the derived data rules for the cyclone tracking workflow in Figure 6.

C.2 Results for MT3D

See Figure 7, 8, 9, 10 for the reasoning results of each sub-workflow. See Figure 11 for the database containing all activated obligations after running the reasoning for all MT3D sub-workflows.

D DATA-GOVERNANCE RULE ENCODING OF CYCLONE TRACKING WORKFLOW

The CMIP6 policy contains multiple rules each may be a link to another nested policy / document.

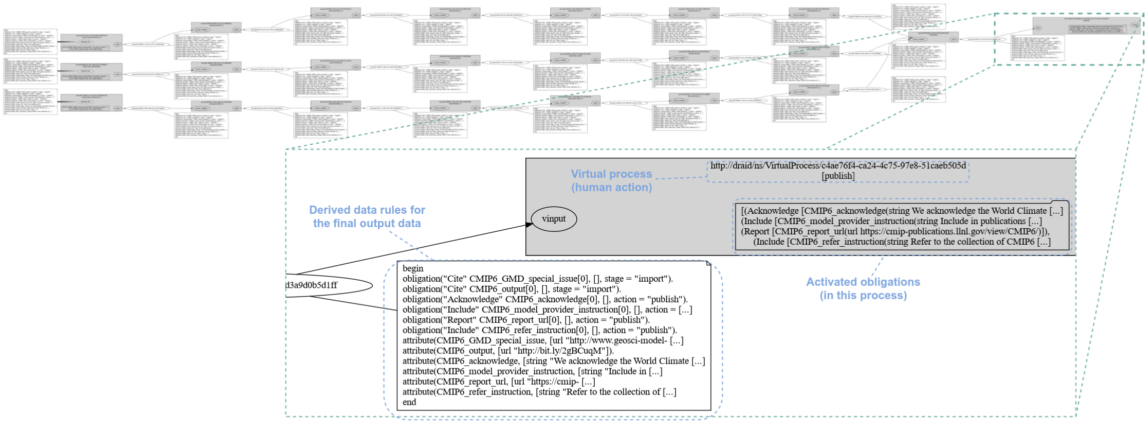

Fig. 6. Derived data rules for cyclone tracking, and the injected virtual process "publish"

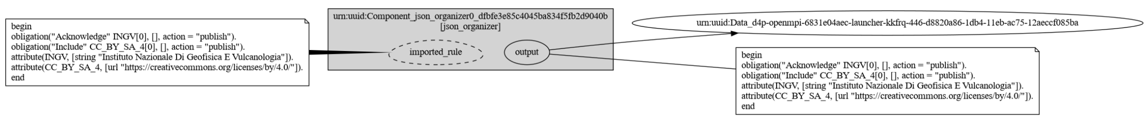

Fig. 7. MT3D reasoning result for create_cmt sub-workflow

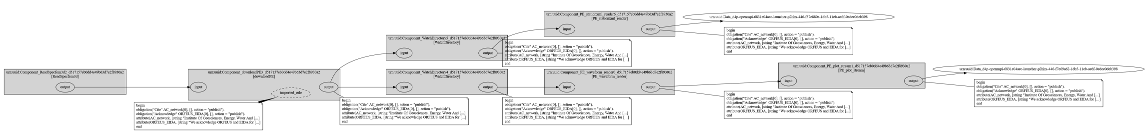

Fig. 8. MT3D reasoning result for download sub-workflow

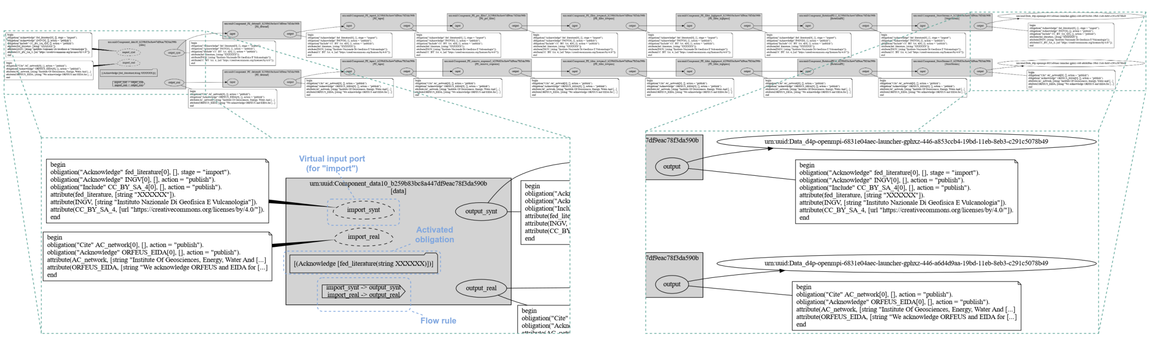

Fig. 9. MT3D reasoning result for preproc sub-workflow

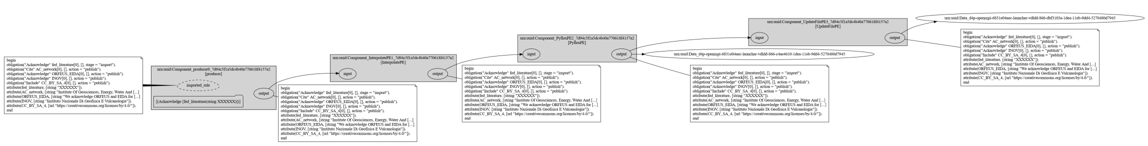

Fig. 10. MT3D reasoning result for pyflex sub-workflow

Persist obligations

Store the obligations to the obligation database, and print all stored obligations.

In [8]:
```
from draid.obligation_store import ObligationStore
import pandas as pd

store = ObligationStore(setting.OBLIGATION_DB)
store.insert(activated_obligations)
store.write()

obs = store.list()
pd.DataFrame(obs, columns=['Triggering process', 'Obligation'])
```

Out[8]:

	Triggering process	Obligation
0	urn:uuid:Component_data10_b259b83bc8a447df9eac...	(Acknowledge [fed_literature(string XXXXXX)])
1	urn:uuid:Component_producer0_7d94c5f1a5dc4b40a...	(Acknowledge [fed_literature(string XXXXXX)])
2	http://draid/ns/VirtualProcess/c37f1d29-effb-4...	(Cite [AC_network(string Institute Of Geoscien...
3	http://draid/ns/VirtualProcess/c37f1d29-effb-4...	(Acknowledge [ORFEUS_EIDA(string We acknowledg...
4	http://draid/ns/VirtualProcess/c37f1d29-effb-4...	(Acknowledge [INGV(string Instituto Nazionale ...
5	http://draid/ns/VirtualProcess/c37f1d29-effb-4...	(Include [CC_BY_SA_4(url https://creativecommo...

Fig. 11. The stored activation conditions emerged from the MT3D reasoning

The document pointed to contains duplicated information, e.g. CMIP6 Data Citation Guidelines[14] , and we discard them. CMIP6 policy contains contextual information, that specifies how its sub-datasets may have different policies. This is automatically addressed by the framework.

When counting the number of sentences, we consider each acknowledge content as one sentence. The sentences in the CMIP6 Data Citation Guidelines are included, because it is defines additional policies on its own, while the other links do not.

Because most rules do not precisely specify when they should be triggered, we must make assumptions based on the context. We believe most of them should be trigger when the user intends to publish the results, therefore we use action = publish as the activation condition. For demonstration purpose, we model some less-strongly implied rules slightly differently: we say that they will trigger when the data is used by the workflow, i.e. stage = import. This may look the same regarding the eventual result at the first glance, but will constitute to different implications. Our reasoning result demonstrates this: they will be triggered multiple times because of the parallel executions.

```
Obligation( Cite CMIP6_GMD_special_issue , [ ], stage = import )
Attribute ( CMIP6_GMD_special_issue, url "http://www.geosci−model−dev.net/special_issue590.html" )

Obligation( Cite CMIP6_output , [ ], stage = import )
```

[14]CMIP6 Data Citation Guidelines: http://bit.ly/2gBCuqM

```
Attribute( CMIP6_output, url "http://bit.ly/2gBCuqM" )

Obligation( Acknowledge CMIP6_acknowledge , [ ], process = publish )
Attribute( CMIP6_acknowledge, "We acknowledge the World Climate Research Programme, which, through its
    ↪ Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modeling
    ↪ groups for producing and making available their model output, the Earth System Grid Federation (ESGF)
    ↪ for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and
    ↪ ESGF." )

Obligation( Include CMIP6_model_provider_instruction , [ ], process = publish )
Attribute( CMIP6_model_provider_instruction, string "Include in publications a table listing the models and
    ↪ institutions that provided model output for research use. In this table and as appropriate in figure legends,
    ↪ use the CMIP6 'official' model names viewable as an html rendering of the CMIP6 source_id controlled
    ↪ vocabulary and an html rendering of institution names recorded in the CMIP6 institution_id controlled
    ↪ vocabulary" )

Obligation( Report CMIP6_report_url , [ ], process = publish )
Attribute( CMIP6_report_url , url "https://cmip-publications.llnl.gov/view/CMIP6/" )

Obligation( Include CMIP6_refer_instruction, [ ], process = publish )
Attribute( CMIP6_refer_instruction, string "Refer to the collection of CMIP6 models as the 'CMIP6 multi-model
    ↪ ensemble' (or similar) and use, as appropriate, phrases like 'CMIP6 multi-model [archive/output/results/
    ↪ simulations/dataset/...]' to describe CMIP6 contributions and products." )
```

E DATA-GOVERNANCE RULE ENCODING OF MT3D WORKFLOW

Here are the rule encodings involved in examining the MT3D workflow. It is split in three parts, each representing a rule origin.

For personal communication. The rule for personal communication is simple, and we directly encode it.

```
Obligation( Acknowledge fed_literature , [ ], stage = import )
Attribute( fed_literature, string "XXXXXX" )
```

For EIDA. The EIDA policy contains nested policy, which refers to additional policies in separate webpages. This nesting requires multiple hops to find the required policies. They are all included in our encoding. In the disclaimer, the EIDA policy specifies some additional rules, e.g. not allowing the user to blame the data provider. They are a mix of contextual information and non-actioning rules. When counting the number of sentences, we include also the ones in the nested policies. But we count only the ones about policies itself, not any more. This is an underestimate of the efforts needed when reading the policies manually.

```
Obligation( Cite AC_network , [ ], action = publish )
```

```
Attribute( AC_network, string "Institute Of Geosciences, Energy, Water And Environment. (2002). Albanian
    ↪ Seismological Network [Data set]. International Federation of Digital Seismograph Networks. https://doi.
    ↪ org/10.7914/SN/AC" )

Obligation( Acknowledge ORFEUS_EIDA , [ ], action = publish )
Attribute( ORFEUS_EIDA, string "We acknowledge ORFEUS and EIDA for providing the waveform data." )
```

For INGV. The data-use policy for INGV contains nested policies of CC-BY. In fact, it says almost nothing more than it is licensed under CC-BY. To avoid duplication, we simply consider CC-BY as a nested policy, and use the Include obligated action to refer to it. When counting the number of sentences, CC-BY is counted as 1 (and 0 implied rules).

```
Obligation( Acknowledge INGV, [ ], action = publish )
Attribute( INGV, string "Instituto Nazionale Di Geofisica E Vulcanologia" )
Obligation( Include CC_BY_4 , [ ], action = publish )
Attribute( CC_BY_4, url "https://creativecommons.org/licenses/by/4.0/" )
```

F ENCODING OF PUBLIC DATA-USE POLICIES

The encodings are presented in each corresponding subsections. Each of them starts with the information and explanation, and then the encoding.

F.1 CC-BY

CC-BY is a widely known licence for shared work. Its URL is: https://creativecommons.org/licenses/by/4.0/.

When counting the number of sentences, we consider the user-facing version, instead of the legal document oriented for interpretation by lawyers.

CC-BY contains two main types of information: what the user is allowed to do and what the user must comply with when doing so (i.e. requirements). The allowed behaviours are all by-default behaviours in our model; the requirements is written as one sentence but contains three distinct actions – 1) crediting the original material and the author, 2) providing a link to the CC-BY licence, and 3) indicate changes made.

We use a simple encoding first (and use this in the table):

```
Attribute( cc_by, str https://creativecommons.org/licenses/by/4.0/ )
Attribute( provider, str Some−Data−Provider, on Original−URL )
Obligation( Acknowledge provider, [ cc_by ], action = publish )
Obligation( ProvideLink cc_by, [ ], action = publish )
Obligation( IndicateChanges provider, [ cc_by ], action = publish )
```

In this encoding, the data provider and data url are both specified within the provider attribute. The 1st obligation statement (with Acknowledge) requirement specifies the crediting action; the 2nd obligation statement (with ProvideLink) specifies the link provide action; the 3rd obligation statement (with IndicateChanges) specifies the last action.

User of this data can change the "provider" attribute through flow rules, and therefore allowing further users to compare changes to this output instead of the original data. This is a possible interpretation to the CC-BY's rule of indicating changes.

But there is a drawback that the original data author gets removed too. To solve this, one can define the provider and link as two different attributes. Another drawback is that this encoding pushes all definition jobs to the framework's core language of obligated actions, etc. It doesn't make use of ontologies to specify obligated action classes or attribute names to facilitate such distributed but interoperable context. Therefore, to illustrate how ontologies are used, we assume CC has a separate namespace cc and specifies the classes or names in it. Therefore, we can do an encoding similar to this:

```
Attribute( cc:cc_by, str https://creativecommons.org/licenses/by/4.0/ )
Attribute( :provider, str Some−Data−Provider )
Attribute( :past_version, url Original−URL )
Obligation( cc:Acknowledge :provider :past_version, [ cc:cc_by ], action = publish )
Obligation( :Include cc:cc_by, [ ], action = publish )
Obligation( cc:IndicateChanges :past_version, [ cc:cc_by ], action = publish )
```

In this way, the Dr.Aid framework author is no longer the sole body who can specify the definitions (for action classes, atrtibute names, etc). In particular, the definition of cc:Acknowledge is different from the default definition provided by the core language of Dr.Aid. The users are able to change the URL without affecting the original provider too.

Again, they are illustrations of several potential ways to encode the policy. We merely exposed the ambiguities within the original policy by formally modelling them, and provide different solutions to them.

F.2 Global CMT Catalogue

The page containing the data-use policy is at: https://www.globalcmt.org/CMTcite.html. This policy has nested policies.

The third rule requires proper citation to the exact rules in the website. The idea solution is to use our language to model the rules for each dataset and associate that directly with the data, and thus removing the need to look up. Our language is able to model them, so we assume they are one rule and is properly modelled.

The fourth rule is about the data from old pre-digital collections. It provides three papers, but did not explain how the user should react. We assume this means the user should properly acknowledge either all of them or the used ones. This is within the capability of our model.

There is an option for doing the first two citations or the third or fourth citation (or all of them) in the original rule.

```
Attribute( CMT_meth_app, str "Dziewonski, A. M., T.−A. Chou and J. H. Woodhouse, Determination of earthquake
    ↪ source parameters from waveform data for studies of global and regional seismicity, J. Geophys. Res., 86,
    ↪ 2825−2852, 1981. doi:10.1029/JB086iB04p02825" )
Obligation( Acknowledge CMT_meth_app, [ ], action = publish )

Attribute( CMT_analysis, str "Ekstrm, G., M. Nettles, and A. M. Dziewonski, The global CMT project 2004−2010:
    ↪ Centroid−moment tensors for 13,017 earthquakes, Phys. Earth Planet. Inter., 200−201, 1−9, 2012. doi
    ↪ :10.1016/j.pepi.2012.04.002" )
```

```
Obligation( Acknowledge CMT_analysis, [ ], action = publish )

Attribute( CMT_study_coll, url "http://www.globalcmt.org/Events/" )
Obligation( Cite CMT_study_coll, [ ], action = publish )

Attribute( CMT_analysis, str "
    Ekstrm, G., and M. Nettles, Calibration of the HGLP seismograph network and centroid−moment tensor analysis
        ↪ of significant earthquakes of 1976, Phys. Earth Planet. Inter., 101, 219−243, 1997. doi:10.1016/S0031
        ↪ −9201(97)00002−2

    Huang, W. C., E. A. Okal, G. Ekstrm, and M. P. Salganik, Centroid moment tensor solutions for deep
        ↪ earthquakes predating the digital era: The World−Wide Standardized Seismograph Network dataset
        ↪ (1962−1976), Phys. Earth Planet. Inter., 99, 121−129, 1997. doi:10.1016/S0031−9201(96)03177−9

    Chen, P. F., M. Nettles, E. A. Okal, and G. Ekstrm, Centroid moment tensor solutions for intermediate−depth
        ↪ earthquakes of the WWSSN−HGLP era (1962−1975), Phys. Earth Planet. Inter., 124, 1−7, 2001. doi
        ↪ :10.1016/S0031−9201(00)00220−X
" )
Obligation( Acknowledge CMT_analysis, [ ], action = publish )
```

F.3 CORDEX

The policy is stated in https://www.hereon.de/imperia/md/assets/clm/cordex_terms_of_use.pdf. This policy has nested policies.

There are different policies for data given to users with different purposes, names research or education or commercial. We model them as three different rules, and different one of them can be attached to the model when distributing the model.

The last rule essentially specifies another acknowledge requirement, but in a less direct way. That requires acknowledging the proper publication associated with the dataset used. This is the direct intention of our framework, so we consider this modelled.

In addition to the normal terms, we added another attribute to represent the scope when the data is still considered as CORDEX (derived) data, and refer to it in all validity bindings. This is optional, and we did this to demonstrate a potential usage of the language and the framework – when a process considers the output is no longer a derivation of CORDEX, it can delete this attribute, and all associated CORDEX obligations are deleted too.

```
Attribute( CORDEX, url, "https://www.hereon.de/imperia/md/assets/clm/cordex_terms_of_use.pdf" )

Obligation( Prohibited, [CORDEX], purpose != research )
Obligation( Prohibited, [CORDEX], purpose != education )
Obligation( Prohibited, [CORDEX], purpose != commercial )
```

```
Attribute( CORDEX_ack, str "We acknowledge the World Climate Research Programme's Working Group on
    ↪ Regional Climate, and the Working Group on Coupled Modelling, former coordinating body of CORDEX
    ↪ and responsible panel for CMIP5. We also thank the climate modelling groups (listed in Table XX of this
    ↪ paper) for producing and making available their model output. We also acknowledge the Earth System Grid
    ↪  Federation infrastructure an international effort led by the U.S. Department of Energy's Program for
    ↪ Climate Model Diagnosis and Intercomparison, the European Network for Earth System Modelling and
    ↪ other partners in the Global Organisation for Earth System Science Portals (GO-ESSP)." )
Obligation( Acknowledge CORDEX_ack, [CORDEX], action = publish )

Attribute( CORDEX_doi, str "I understand that Digital Object Identifiers (DOI's used, for example, in journal
    ↪ citations) together with a citation reference will be assigned to some of the CORDEX datasets during the
    ↪ DataCite data publication process, and when available and as appropriate, I will cite CORDEX data by
    ↪ these citation references in my publications. I will consult the CORDEX data website (http://cordex.dmi.dk)
    ↪  to learn how to do this." )
Obligation( Include CORDEX_doi, [CORDEX], action = publish )
```

F.4 ISMD

```
Attribute( ISMD_ack, str "Marco Massa, Ezio DAlema, Sara Lovati, Simona Carannante, Gianlorenzo Franceschina,
    ↪ Paolo Augliera (2016). INGV Strong Motion Data (ISMD) v2.1, Istituto Nazionale di Geofisica e
    ↪ Vulcanologia (INGV). https://doi.org/10.13127/ismd.2.1" )
Obligation( Acknowledge ISMD_ack, [ ], action = publish )
```

F.5 RCMT

The policy is stated directly on http://rcmt2.bo.ingv.it/, which has nested policies. The data is licensed under CC-BY.

It also synthesizes data from several different sources, each has their own policies with the acknowledgment requirement. We stop here, as this policy did not indicate that the user should also provide acknowledgment to them.

```
Attribute( RCMT_ack, str "Pondrelli, S. (2002). European-Mediterranean Regional Centroid-Moment Tensors
    ↪ Catalog (RCMT) [Data set]. Istituto Nazionale di Geofisica e Vulcanologia (INGV). https://doi.org/10.13127/
    ↪ rcmt/euromed" )
Obligation( Acknowledge RCMT_ack, [ ], action = publish )
Obligation( IndicateChanges, [ ], action = publish )
```

F.6 MIMIC

The policy is stated in this page https://mimic.physionet.org/about/acknowledgments/, and some additional information are in https://mimic.physionet.org/gettingstarted/access/.

This repository contains rules for two types of assets, the MIMIC data and the MIMIC code. Different rules apply to them.

F.6.1 Data.

```
Attribute( MIMIC_ack, str "MIMIC−III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L,
    ↪ Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI:
    ↪  10.1038/sdata.2016.35. Available at: http://www.nature.com/articles/sdata201635" )
Obligation( Acknowledge MIMIC_ack, [ ], action = publish )

Attribute( MIMIC_data, str "Pollard, T. J. & Johnson, A. E. W. The MIMIC−III Clinical Database http://dx.doi.org
    ↪ /10.13026/C2XW26 (2016)." )
Obligation( Acknowledge MIMIC_data, [ ], action = publish )

Attribute( PhysioNet_ack, str "Physiobank, physiotoolkit, and physionet components of a new research resource for
    ↪ complex physiologic signals. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov P, Mark RG,
    ↪ Mietus JE, Moody GB, Peng C, and Stanley HE. Circulation. 101(23), pe215e220. 2000." )
Obligation( Acknowledge PhysioNet_ack, [ ], action = publish )
```

F.6.2 Code.

```
Attribute( MIMIC_code, str "Johnson, Alistair EW, David J. Stone, Leo A. Celi, and Tom J. Pollard. "The MIMIC Code
    ↪ Repository: enabling reproducibility in critical care research." Journal of the American Medical Informatics
    ↪ Association (2017): ocx084." )
Obligation( Acknowledge MIMIC_code, [ ], action = publish )
```

F.7 CPRD

The post-use policy is stated for each dataset on https://www.cprd.com/DOIs. There are multiple datasets each with their own DOIs. We use one of the real datasets in the example encoding, because the synthetic datasets contains fewer rules.

In addition, accessing their data requires application by going through https://www.cprd.com/data-access where additional policies are stated in the application form.

The special part of it is that the data and results shall be kept confidential and used only by the applicant, which is what the Prohibited obligations state. But this can be lifted under certain conditions, which can be expressed as a process removing the CPRD_controlled attribute (thus removing the bound obligations).

```
Attribute( CPRD_gold_mar, str Citation: Clinical Practice Research Datalink. (2021). CPRD GOLD March 2021 (
    ↪ Version 2021.03.001) [Data set]. Clinical Practice Research Datalink. https://doi.org/10.48329/WH2F−8168 )
Obligation( Acknowledge CPRD_gold_mar, [ ], action = publish )

Attribute( CPRD_controlled, url https://www.cprd.com/Data−access )
Obligation( Prohibited, [CPRD_controlled], action = publish )
Obligation( Prohibited, [CPRD_controlled], user != SomeUserId )
```

F.8 PIMA

This dataset is licensed under CC-0, but proper acknowledgement is encouraged. This page contains relevant information: https://www.kaggle.com/uciml/pima-indians-diabetes-database.

```
Attribute( PIMA_ack, str "Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using
    ↪ the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium
    ↪ on Computer Applications and Medical Care (pp. 261−−265). IEEE Computer Society Press." )
Obligation( Acknowledge PIMA_ack, [ ], action = publish )
```

F.9 ISC

There are multiple sub-datasets contained in this data source. The collective policy is accessible through http://www.isc.ac.uk/citations/.

This policy contains nested policies for different sub-items. Each of them has different specific policies, but the general form is to properly acknowledge the dataset and the research work being used. Therefore, we use the first one of them, ISC Bulletin, as the encoding example.

```
Attribute( ISC_product, str "International Seismological Centre (20XX), On−line Bulletin, https://doi.org/10.31905/
    ↪ D808B830" )
Obligation( Acknowledge ISC_product, [ ], action = publish )

Attribute( ISC_art_a, str "Bondr, I. and D.A. Storchak (2011). Improved location procedures at the International
    ↪ Seismological Centre, Geophys. J. Int., 186, 1220−1244, doi: 10.1111/j.1365−246X.2011.05107.x" )
Obligation( Acknowledge ISC_art_a, [ ], action = publish )

Attribute( ISC_art_b1, str "Storchak, D.A., Harris, J., Brown, L., Lieser, K., Shumba, B., Verney, R., Di Giacomo, D.,
    ↪ Korger, E. I. M. (2017). Rebuild of the Bulletin of the International Seismological Centre (ISC), part 1: 1964
    ↪ 1979. Geosci. Lett. (2017) 4: 32. doi: 10.1186/s40562−017−0098−z" )
Obligation( Acknowledge ISC_art_b1, [ ], action = publish )

Attribute( ISC_art_b2, str "
Storchak, D.A., Harris, J., Brown, L., Lieser, K., Shumba, B., Di Giacomo, D. (2020) Rebuild of the Bulletin of the
    ↪ International Seismological Centre (ISC)part 2: 19802010. Geosci. Lett. 7: 18, https://doi.org/10.1186/s40562
    ↪ −020−00164−6" )
Obligation( Acknowledge ISC_art_b2, [ ], action = publish )

Attribute( ISC_art_c, str "R J Willemann, D A Storchak (2001). Data Collection at the International Seismological
    ↪ Centre, Seis. Res. Lett., 72,, 440−453, doi: https://doi.org/10.1785/gssrl.72.4.440" )
Obligation( Acknowledge ISC_art_c, [ ], action = publish )

Attribute( ISC_art_d, str "Di Giacomo, D., and D.A. Storchak (2016). A scheme to set preferred magnitudes in the ISC
    ↪  Bulletin, J. Seism., 20(2), 555−567, doi: 10.1007/s10950−015−9543−7" )
```

```
Obligation( Acknowledge ISC_art_d, [ ], action = publish )

Attribute( ISC_art_e1, str "Lentas, K., Di Giacomo, D., Harris, J., and Storchak, D. A. (2019). The ISC Bulletin as a
    ↪ comprehensive source of earthquake source mechanisms, Earth Syst. Sci. Data, 11, 565−578, doi: https://doi.
    ↪ org/10.5194/essd−11−565−2019" )
Obligation( Acknowledge ISC_art_e1, [ ], action = publish )

Attribute( ISC_art_e2, str "Lentas, K. (2018). Towards routine determination of focal mechanisms obtained from first
    ↪ motion P−wave arrivals, Geophys. J. Int., 212(3), 16651686. doi: 10.1093/gji/ggx503" )
Obligation( Acknowledge ISC_art_e2, [ ], action = publish )

Attribute( ISC_art_f, str "Adams, R.D., Hughes, A.A., and McGregor, D.M. (1982). Analysis procedures at the
    ↪ International Seismological Centre. Phys. Earth Planet. Inter. 30: 85−93, doi: https://doi.org
    ↪ /10.1016/0031−9201(82)90093−0" )
Obligation( Acknowledge ISC_art_f, [ ], action = publish )
```

F.10 IRIS

This data source also contains diverse data and therefore diverse rules, stated on https://www.iris.edu/hq/iris_citations. It also has nested policies to FSDN.

This policy set contains different policies for different assets. Most rules are simply requiring the user to properly acknowledge the data being used.

The second rule is about properly acknowledging FDSN object. This is the same as for EIDA data. Therefore, for simplicity, we treat this as one rule in this example, and use the Cite obligated action.

```
Attribute( IRIS_report, url "https://www.iris.edu/hq/forms/submit_citation" )
Obligation( Report IRIS_report, [ ], action = publish )

Attribute( IRIS_service, str "The facilities of IRIS Data Services, and specifically the IRIS Data Management Center,
    ↪ were used for access to waveforms, related metadata, and/or derived products used in this study. IRIS Data
    ↪ Services are funded through the Seismological Facilities for the Advancement of Geoscience (SAGE) Award
    ↪  of the National Science Foundation under Cooperative Support Agreement EAR−1851048." )
Obligation( Acknowledge IRIS_service, [ ], action = publish )

Attribute( IRIS_FDSN, url "https://www.fdsn.org/networks/citation/" )
Obligation( Cite IRIS_FDSN, [ ], action = publish )

Attribute( IRIS_GSN, str "Global Seismographic Network (GSN) is a cooperative scientific facility operated jointly by
    ↪  the Incorporated Research Institutions for Seismology (IRIS), the United States Geological Survey (USGS),
    ↪ and the Seismological Facilities for the Advancement of Geoscience (SAGE) Award of the National Science
    ↪ Foundation (NSF), under Cooperative Support Agreement EAR−1851048." )
```

```
Obligation( Acknowledge IRIS_GSN, [ ], action = publish )

Attribute( IRIS_PASSCAL_Polar, str "Acknowledgment – In any publications or reports resulting from the using IRIS
    ↪ ' Polar–specific instruments or support, please include the following statement in the acknowledgment
    ↪ section. You are also encouraged to acknowledge NSF and IRIS in any contacts with the news media or in
    ↪ general articles.\nThe seismic instruments were provided by the Incorporated Research Institutions for
    ↪ Seismology (IRIS) through the PASSCAL Polar Support Services. Data collected will be available through
    ↪ the IRIS Data Management Center. The facilities of the IRIS Consortium are supported by the National
    ↪ Science Foundations Seismological Facilities for the Advancement of Geoscience (SAGE) Award under
    ↪ Cooperative Support Agreement OPP–1851037." )
Obligation( Include IRIS_PASSCAL_Polar, [ ], action = publish )

Attribute( IRIS_Trans, str "Data from the TA network were made freely available as part of the EarthScope USArray
    ↪ facility, operated by Incorporated Research Institutions for Seismology (IRIS) and supported by the
    ↪ National Science Foundation, under Cooperative Agreements EAR–1261681." )
Obligation( Acknowledge IRIS_Trans, [ ], action = publish )

Attribute( IRIS_PASSCAL_Mag, str "The magnetotelluric instruments were provided by the Incorporated Research
    ↪ Institutions for Seismology (IRIS) through the PASSCAL Instrument Center at New Mexico Tech. Data
    ↪ collected will be available through the IRIS Data Management Center. The facilities of the IRIS Consortium
    ↪ are supported by the National Science Foundations Seismological Facilities for the Advancement of
    ↪ Geoscience (SAGE) Award under Cooperative Support Agreement EAR–1851048." )
Obligation( Acknowledge IRIS_PASSCAL_Mag, [ ], action = publish )

Attribute( IRIS_Edu, str "Materials provided by the IRIS Education and Public Outreach Program have been used in
    ↪ this study. The facilities of the IRIS Consortium are supported by the National Science Foundations
    ↪ Seismological Facilities for the Advancement of Geoscience (SAGE) Award under Cooperative Support
    ↪ Agreement EAR–1851048." )
Obligation( Acknowledge IRIS_Edu, [ ], action = publish )

Attribute( IRIS_OBSIC, str "Data used in this research were provided by instruments from the Ocean Bottom
    ↪ Seismograph Instrument Center (obsic.who.edu) which is funded by the National Science Foundation.
    ↪ OBSIC data are archived at the IRIS Data Management Center ([url=http://www.iris.edu]http://www.iris.
    ↪ edu[/url]) which is funded by the National Science Foundations Seismological Facilities for the
    ↪ Advancement of Geoscience (SAGE) Award under Cooperative Support Agreement EAR–1851048." )
Obligation( Acknowledge IRIS_OBSIC, [ ], action = publish )
```

F.11 OGL: Open Government Licence

This licence is stated on https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/. This is a general licence and each dataset may specify their own acknowledgment statement.

```
Attribute( OGL_ack, str "Contains public sector information licensed under the Open Government Licence v3.0." )
Obligation( Acknowledge OGL_ack, [ ], action = publish )
```

F.12 World Bank

This policy is stated in https://www.worldbank.org/en/about/legal/terms-of-use-for-datasets. It contains nested policies, which refer to CC-BY and potential separate policies in its 3rd-party data. It explicitly re-specifies several aspects of CC-BY, so they are counted as a part of the policy.

Maybe because this policy is more close to the legal document, there is a large amount of disclaimer and contextual information. They constitute the general form of rules, but they are normally not actioning rules.

```
Obligation( Acknowledge WB, [ ], process = "publish" )
Attribute( WB, string "The World Bank: Dataset name: Data source (if known)" )
Obligation( Include CC_BY_SA_4 , [ ], process = "publish" )
Attribute( CC_BY_SA_4, url "https://creativecommons.org/licenses/by/4.0/" )

Obligation( Include WB_communicate, [ ], null)
Attribute( WB_communicate, str If you have questions, seek to use Datasets on license terms other than the ones
    ↪ described above, or wish to make other comments, please contact us at +1 202 473 7824 or +1 800 590 1906,
    ↪ or by sending an email to data@worldbank.org. )
```