Dr.Aid: an automated formal framework to support data-governance rule compliance for decentralized collaboration

Rui Zhao

Supervised by: Petros Papapanagiotou, Malcolm Atkinson, Jacques Fleuriot

University of Edinburgh

2021-05-24

Rules – heterogeneity, similarity and redundancy

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

Problem to solve

- Large collection of ... data
    - personal
    - sensitive
    - valuable

Problem to solve

- Large collection of ... data
  - personal
  - sensitive
  - valuable
- Issues of ...
  - data privacy
  - ownership
  - governance

Problem to solve

- Large collection of ... data
  - personal
  - sensitive
  - valuable
- Issues of ...
  - data privacy
  - ownership
  - governance

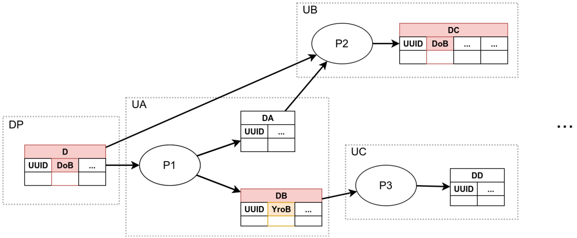
Compliance checking remains ...

- manual
- time consuming
- error prone

Example use case

Data governance rules associated with dataset D by data provider DP

Users should properly acknowledge the data provider in their publications, in a form similar to “This work ...”.
Field DoB (Date of Birth) in the data is potentially *sensitive* and any use of it should be reported to the data provider, through the URL report.example.ac.



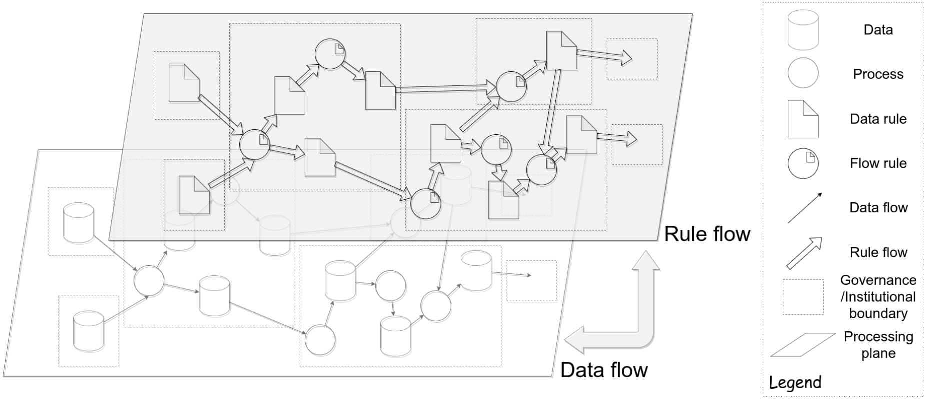
Issues identified

1. (Personnel) **Scattering**: data processing is *multi-institutional* (providers don't work together with users);
2. (Rule) **Propagation**: derived data can be used as input data further;
3. (Rule) **Diversity**: policies can be more than access control, e.g. *obligations*;
4. **Dynamic** (rule) **application**: processes can change the policies applied to data;
5. (Rule) **Combination** and **separation**: processes are multi-input-multi-output (MIMO).

Related research

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)	✓	✓	✓	✓	✓	✓

From **data flow** to **rule flow**



Rule flow

Data flow

Data

Process

Data rule

Flow rule

Data flow

Rule flow

Governance /Institutional boundary

Processing plane

Legend

What is Dr.Aid

Dr.Aid (Data Rule Aid) is a *logic-based* framework for *automated compliance checking* of data governance rules over data flow *graphs*.

What is Dr.Aid

Dr.Aid (Data Rule Aid) is a *logic-based* framework for *automated compliance checking* of data governance rules over data flow *graphs*.

Highlights:

- Logic-based (situation calculus)
- All 5 issues addressed, in particular:
  - Recognises rule diversity
  - Addresses dynamic rule application
  - Supports MIMO

Model

- Language:
  - **data rule**, for data-governance rules for multi-staged processing, e.g. "*users must properly acknowledge the data providers*"
  - **flow rule**, for the *changes* of data rules in each *process* as a result of data transportation and transformation, e.g. "*column 3 from input 1 is changed to column 2 on output 2*"
- Supports *obligation*: user must do something after using the data
- Retrospective analysis from provenance
  - CWLProv (file-oriented)
  - S-Prov (data-streaming)

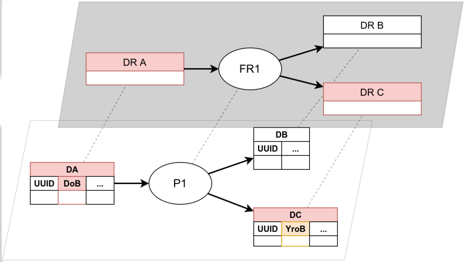
Language – example

Data-use policy (as *data rule*)

Field DoB (Date of Birth) in the data is potentially *sensitive* and any use of it should be reported to the data provider.

Process information (as *flow rule*)

- changes column DoB from input 1 to column YroB on output 2;
- removes column DoB from input 1 on output 1;
- propagates the rest of the data unaffected regarding the data-use policy.



Language – example modelled

This dataset contains potentially sensitive information in column DoB.
Any use of it should be reported to the data providers through link report.example.ac.

Natural-language policy

attribute(pf, column "DoB")

Name Type Value

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

Name Type Value

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

Obligated action Action argument Validity binding Activation condition

User notation

attribute(pf, column, "DoB", [input1, pf_1], s0)

History (ID) Situation

attribute(ru, url, "report.example.ac", [input1, ru_1], s0)

History (ID) Situation

obligation(report, [[input1, ru_1]],
[[input1, pf_1]], "action = *", input1, s0)

Port Situation

Situation calculus representation

obligation(report attribute(ru, url "report.example.ac"), [attribute(pf, column "DoB")], action = *)

Nested equivalent view

Language – example modelled and query

Querying is analogous to the projection task, i.e. querying predicates (fluents) that hold at the targeted final situation:

$$\begin{aligned} S_f = Do(&pr(input1, [output1, output2]) : \\ &delete(input1, output1, *, column, "DoB") : \\ &edit(input1, output2, *, column, "DoB", column, "YroB") : \\ &end([output1, output2]), s0). \end{aligned}$$

$$? - attribute(N, T, V, X, S_f).$$

Language – example modelled and query

Querying is analogous to the projection task, i.e. querying predicates (fluents) that hold at the targeted final situation:

$$\begin{aligned} S_f = Do(pr(&input1, [output1, output2]) : \\ & delete(input1, output1, *, column, ''DoB'') : \\ & edit(input1, output2, *, column, ''DoB'', column, ''YroB'') : \\ & end([output1, output2]), s0). \end{aligned}$$

$$?-attribute(N, T, V, X, S_f).$$

resulting in

$$\begin{aligned} & attribute(ru, url, ''report@example.ac'', [output1, input1, pf_1], S_f) \\ & attribute(pf, column, ''YroB'', [output2, input1, pf_1], S_f) \\ & attribute(ru, url, ''report@example.ac'', [output2, input1, pf_1], S_f) \end{aligned}$$

... and the same procedure for *obligations*.

Language – Data rule

- *attributes* (N, T, V), describing properties of the data
  - a name N
  - a type T
  - a value V
- *obligations* (OD, VB, AC)
  - an obligation definition OD (the obligated action to perform upon activation)
  - a validity binding VB (describing additional applicability constraints)
  - an activation condition AC (the triggering condition)

Language – flow rule

- **propagate** $pr(P_{in}, Ps_{out})$
- **edit** $edit(P_{in}, P_{out}, N, T, V, T_{new}, V_{new})$
- **delete** $delete(P_{in}, P_{out}, N, T, V)$

Language – flow rule

- **propagate** $pr(P_{in}, Ps_{out})$
- **edit** $edit(P_{in}, P_{out}, N, T, V, T_{new}, V_{new})$
- **delete** $delete(P_{in}, P_{out}, N, T, V)$

$$pr(input1, [output1, output2])$$
$$delete(input1, output1, *, column, DoB)$$
$$edit(input1, output2, *, column, DoB, column, YroB)$$

$*$ represents *anything*, used as wildcard.

Situation calculus formalisation

Situation calculus models things into three components:

- *fluent*: predicate with *situation*, holding different truth values at different *situation*s
- *situation*: state of the world, where the initial situation specifies the initial state of the world
- *action*: behaviour to change the world, applying to a *situation* and resulting in a new *situation*

Situation calculus formalisation

Situation calculus models things into three components:

- *fluent*: predicate with *situation*, holding different truth values at different *situation*s
- *situation*: state of the world, where the initial situation specifies the initial state of the world
- *action*: behaviour to change the world, applying to a *situation* and resulting in a new *situation*

We model:

- data rule terms as *fluents*
- the initial data rules are specified for the initial *situation*
- the flow rules are modelled as *actions*

Situation calculus formalisation

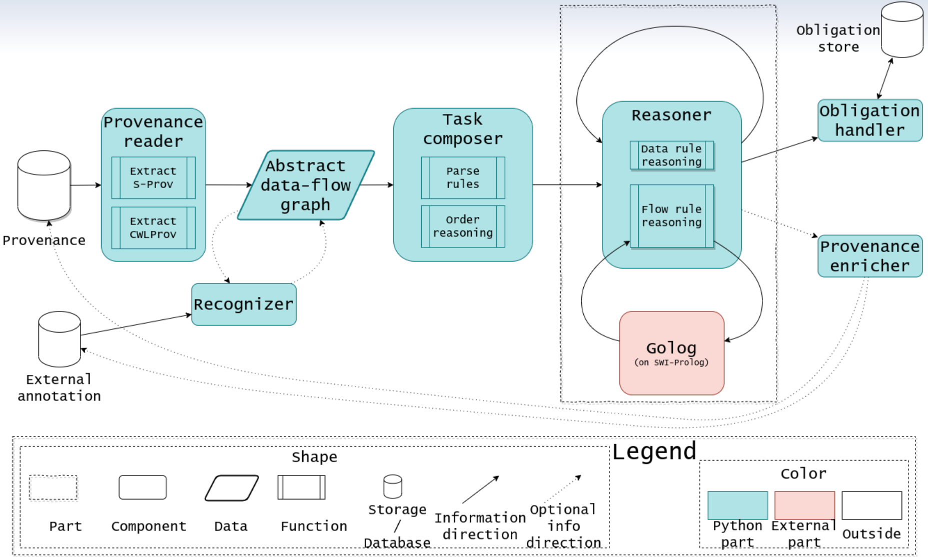
We model:

- data rule terms as *fluents*
- the initial data rules are specified for the initial *situation*
- the flow rules are modelled as *actions*

We do:

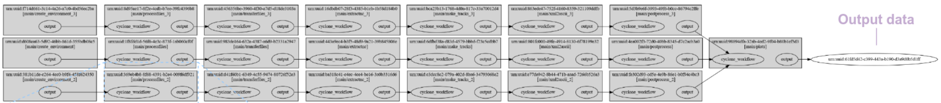
- flow rules as action sequence
- final situation S_f as the situation after performing all flow rules starting from the initial situation s_0
- query all $\mathit{attribute}()$ and $\mathit{obligation}()$ which hold in the final situation S_f

System structure

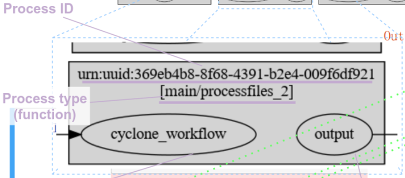


Outcomes from system

Out[5]:



urn:uuid:f714d661-3c14-4a26-a7c0-4bd56ec2ba
[main/create_environment_3]
urn:uuid:8d95ae17-8f2e-4cdb-b7ee-39fc4396b8
[main/processfiles_3]
urn:uuid:456356be-3960-4f30-a7d5-d18de3103a
[main/transferfiles_3]
urn:uuid:16dbdb07-28f3-4385-b1eb-1b58d184b0
[main/extractnc_3]
urn:uuid:bca23b13-1768-4d0a-817c-33a70012d4
[main/make_tracks_3]
urn:uuid:863ede47-752f-4160-8359-321109ddfb
[main/xml2ascii_3]
urn:uuid:5d3b9e6f-3993-49f6-b0cc-86794c2ffe
[main/postprocess_3]
urn:uuid:d608ea63-5d92-46bb-861d-355bdb08c5
[main/create_environment]
urn:uuid:1f8fdd1d-56f6-4e3c-873f-1ab060cf0f
[main/processfiles]
urn:uuid:983de164-632a-4387-a6d9-b2331a2947
[main/transferfiles]
urn:uuid:443a9ec4-b3f5-48d9-9a21-39b845806e
[main/extractnc]
urn:uuid:6dfbd38a-d83d-4579-bbbd-f23c5cdbb7
[main/make_tracks]
urn:uuid:801fc000-49fe-4914-8130-6f78199e32
[main/xml2ascii]
urn:uuid:4ea002f3-72d0-40bb-8345-d7c2acb3a0
[main/postprocess]
urn:uuid:98894dfa-32ab-4ed2-9f04-b80b1ef5d3
[main/plots]
urn:uuid:61fd5d42-c399-443a-b190-d3a9d0b5d1ff
urn:uuid:381b11de-c264-4ec0-b9f4-458f624350
[main/create_environment_2]
urn:uuid:369eb4b8-8f68-4391-b2e4-009f6df921
[main/processfiles_2]
urn:uuid:d41f60b1-6349-4c55-9474-80726f52e3
[main/transferfiles_2]
urn:uuid:ba318c41-c4ec-4ce4-be16-3a0b3316d6
[main/extractnc_2]
urn:uuid:c3dcc8e2-079a-402d-8ba6-34793068e2
[main/make_tracks_2]
urn:uuid:e77de942-8b44-471b-aaad-7266b526a3
[main/xml2ascii_2]
urn:uuid:fe302d93-cd5e-4e9b-86e1-b0f94c4bc3
[main/postprocess_2]
cyclone_workflow
output
Output data



Out[7]:

	Triggering process	Obligation
0	urn:uuid:d608ea63-5d92-46bb-861d-355bdb08c587	(Cite [CMIP6_GMD_special_issue(url http://www....
1	urn:uuid:d608ea63-5d92-46bb-861d-355bdb08c587	(Cite [CMIP6_output(url http://bit.ly/2gBCuqM)])
2	urn:uuid:f714d661-3c14-4a26-a7c0-4bd56ec2ba0d	(Cite [CMIP6_GMD_special_issue(url http://www....
3	urn:uuid:f714d661-3c14-4a26-a7c0-4bd56ec2ba0d	(Cite [CMIP6_output(url http://bit.ly/2gBCuqM)])
4	urn:uuid:381b11de-c264-4ec0-b9f4-458f624350e9	(Cite [CMIP6_GMD_special_issue(url http://www....
5	urn:uuid:381b11de-c264-4ec0-b9f4-458f624350e9	(Cite [CMIP6_output(url http://bit.ly/2gBCuqM)])
6	http://draid/ns/VirtualProcess/ab869295-9f77-4...	(Acknowledge [CMIP6_acknowledge(string We ackn...
7	http://draid/ns/VirtualProcess/ab869295-9f77-4...	(Include [CMIP6_model_provider_instruction(str...
8	http://draid/ns/VirtualProcess/ab869295-9f77-4...	(Report [CMIP6_report_url(url https://cmip-pub...
9	http://draid/ns/VirtualProcess/ab869295-9f77-4...	(Include [CMIP6_refer_instruction(string Refer...

```
Finished Initialization
```

Out[6]:

```
{rdflib.term.URIRef('urn:uuid:d608ea63
ecial_issue590.html)]),
  (Cite [CMIP6_output(url http://bit.l
 rdflib.term.URIRef('urn:uuid:f714d661
ecial_issue590.html)]),
  (Cite [CMIP6_output(url http://bit.l
 rdflib.term.URIRef('urn:uuid:381b11de
ecial_issue590.html)]),
  (Cite [CMIP6_output(url http://bit.ly/2gBCuqM)]),
 rdflib.term.URIRef('http://draid/ns/VirtualProcess/ab869295-9f77-49ed-b99d-3f6f7ca7353a'): [(Acknowledge [CMIP6_acknowledge(string We ackno
wledge 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.)]),
  (Include [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 rend
ering of the CMIP6 source_id controlled vocabulary and an html rendering of institution names recorded in the CMIP6 institution_id controlle
d vocabulary)]),
```

Language coverage

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%

Thanks
for listening!