Modelling Data Rules: Inferenceable Terms and Conditions

Rui ZHAO

Centre for Intelligent Systems and their Applications, School of Informatics, University of Edinburgh

Motivation

Scientific research is becoming increasingly data intensive, and researchers use data from different sources. Different data providers set different usage policies.
In order not to break rules, researchers need to find and read every policy. Data owners don't trust researchers to obey they rules.

More importantly: people often forget!

Real case

IRIS (Incorporated Research Institutions for Seismology) is a US institution providing seismic sensor data. Accessing IRIS data should use their API. When used in class as coursework, multiple simutaneous accesses happened, and that was seen as a DoS attack by IRIS. However, when negotiating, IRIS also doesn't allow us to cache their data locally, because they want to keep track of the number of times their data is used.

— Ian Main, School of GeoSciences

Future Vision

Researchers

1. Pick necessary existing workflow components, and develop relevant new processes where needed
2. Compose the workflow / processes with selected data (input)
3. Execute the workflow, and obtain result
4. Review result data
   - If useful, end the experienment
   - If not useful (yet), restart

We'd expect the infrastructure to

1. handle necessary rountine steps for the researcher (e.g. download data)
2. keep track of compliance and changes to policies before, during and after processing
3. warn the researcher, if any of the steps may break the policy of a dataset

Abstraction & Assumption

- Rules associated with input may apply to the output of a process or be dealt with by the process
- The data processing can be modeled as a computational graph (e.g. workflow)
- Human interaction is modelled as a process
- We build our system on top of *provenance data*, the execution trace of a workflow
  - It abstracts the underlying heterogeneity
  - Both retrospective and prospective provenance are acceptable

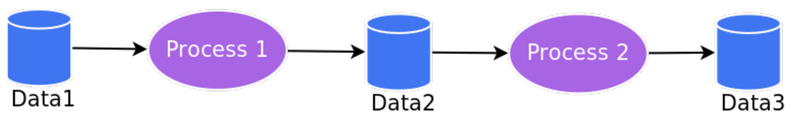
What policy should I obey?

You are a researcher who works with data supplied with obligations

- Each data source you use has its own policies which you intend to honour
- Your work is modeled as a series of computational steps
  - Each step may produce a dataset and/or streams the output(s) to other steps

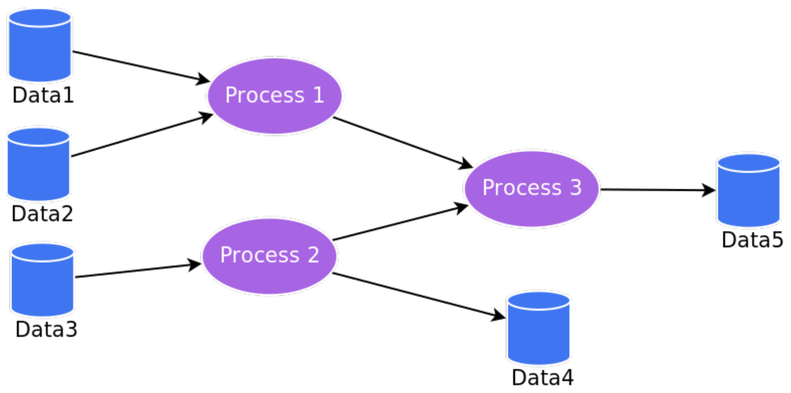
You want to know for each of your produced data, what policies you should follow

Simple case



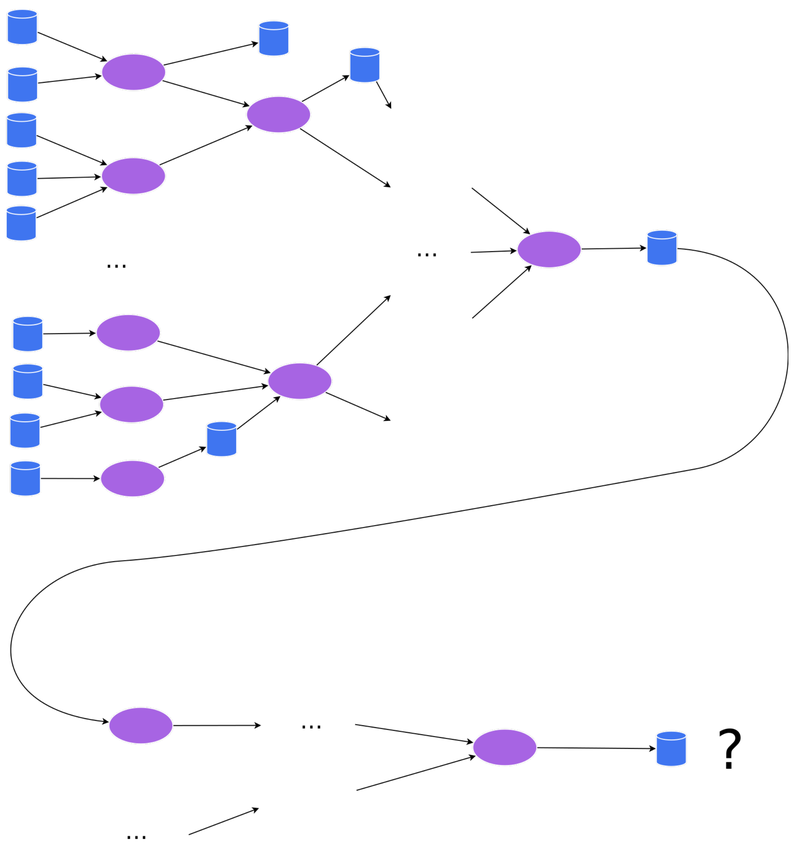
Easy, right?

Case you think you have



Still easy?

Real-world case you may be actually in



Whaaaaaat?

Objectives

- Develop a **formal model** to *represent the data policies* and reasoning
  - A representation of the properties which the rules may refer to
  - A representation of rule propagation through processes
- Prototype mechanisms
  - to model rules
  - to associate rules with data
  - to infer the behaviour of a workflow
  - to assess whether a rule has broken
- Deliver an **inference system** using these
- Collect **example rules** (from real use cases) that demonstrates the capability

Methodology

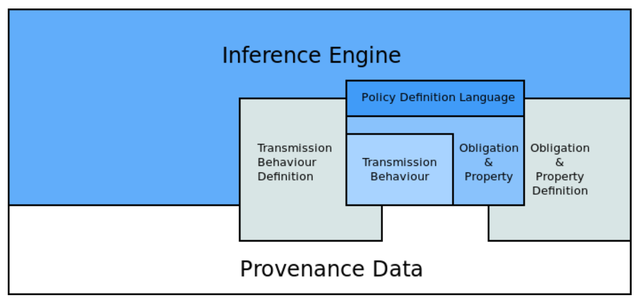


Figure 1:System view

- Extra annotations are added to the metadata (linked in provenance) to serve as the input of inference
  - Each *dataset* is associated with a set of ***obligations*** and their cooresponding ***properties***
  - Each *processing element* has its associated ***transmission behaviour***
- *Obligation* and *property* are transmitted through the workflow
- Each *processing element* produces *obligation* and *property* according to its *transmission behaviour* and the *obligation* and *property* it takes as input

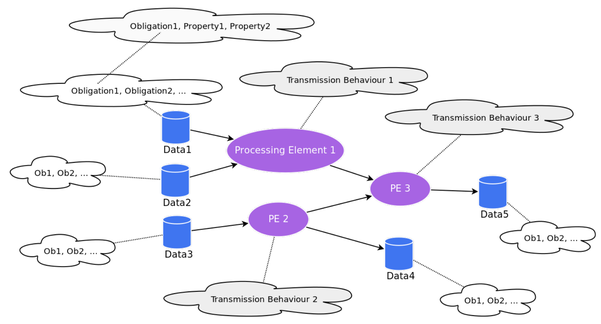


Figure 2:Workflow view

Ontologization

We will develop a corresponding *ontology* for *obligations* and *transmission behaviours*

- It allows extension
- Human-readable forms can be defined
- A universal exchanging form
- Datasources can attach their rules to their data

Contact Information

- Rui Zhao: rui.zhao@inf.ed.ac.uk
- Supervised by
  - Malcolm Atkinson: mpa@staffmail.ed.ac.uk
  - Jacques Fleuriot: jdf@inf.ed.ac.uk

cisa | Centre for Intelligent Systems and their Applications

THE UNIVERSITY of EDINBURGH