TDR: What does partial compliance mean?

Markus.Buchhorn@anu.edu.au

Why me?

- Ex-astronomer
  - *Really* picky about the scientific method, metrics, measurements & metadata
- Multiple hats
  - ANU, APAC, GrangeNet, and participation in many programs
  - Lots of use-cases, in a broad diversity of disciplines
    - Physical sciences, social sciences, education and research
    - Scholarly input, as well as scholarly outputs
    - Small to large scale, short to long term
    - All of it extremely valuable
- APAC/APSR survey of e-research collections
  - Around 50 projects analysed in-depth
- Really keen on the idea of ‘certification’ and ‘recognition’

Disclaimer

- Asked by APSR/NLA to give this talk
  - Suggested I be ‘contentious’… ;-)
- Have not lived the experience like some here
  - Not looking to be editor
    - Though I did spot a few grammatical errors
  - Have not read the draft repeatedly
    - Keep finding new angles
    - Have missed some things, and the updated thinking

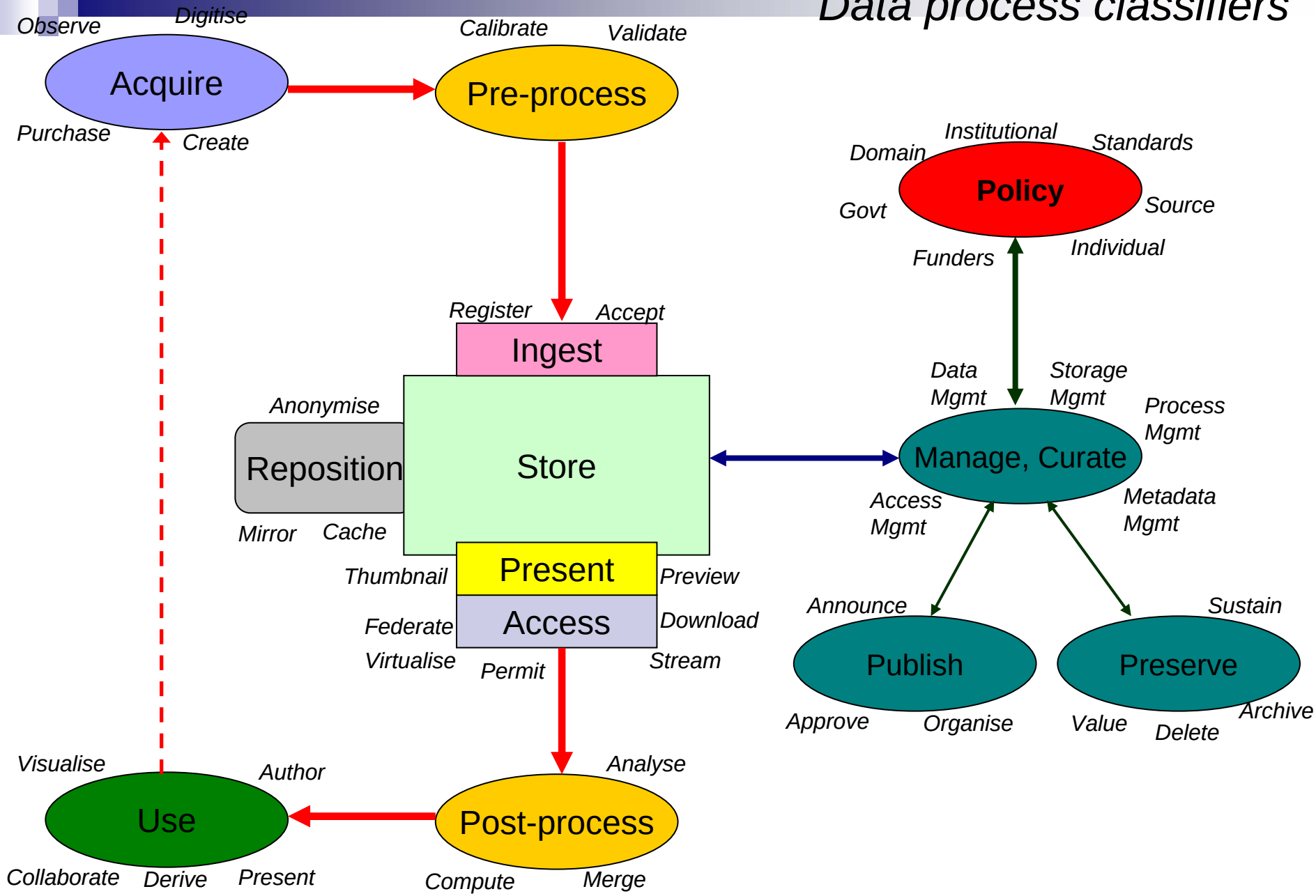
What does *"what does partial compliance mean"* mean?

- What does "partial compliance" mean?
  - Measurements, metrics and methods
- What does partial compliance "mean"?
  - i.e. who cares, and why?

What is “trust”?

- Broad review
  - Philosophy, sociology, dictionaries, …
    - Not by me!
- Boiled down to:
  - Makes life *predictable*
  - Creates a *community*
  - Makes it *easier to collaborate*

Data process classifiers



What does “partial compliance” mean?

- It is a measure
  - Can you only be “in” or “out”?
  - Can you be “some number” along the path?
    - On a scale of 0-100, you’re a …
- Is “compliance” like “pregnancy”?
  - Yes: Getting there is half the work…
  - Yes: You can be or be not…
  - No: You can go backwards, and sideways
    - Staying compliant…

Measurements and metrics

- Can you measure a degree of compliance
  - Per item, per category, overall?
  - Currently: Thought about it, wrote it down, built it, tested it
    - These are steps on a path,
    - but it's the quality of the implementation we're measuring
- Can we associate some quantitative measurements of progress?
- Can we compare the impact of individual compliance elements against each other?
  - "this element is twice as important as that one"
    - "they're all equally important"
  - "this repository is twice as compliant as that one"
  - Probably not…
    - It may depend on who is measuring

Measurements and metrics

- Policies: what you'd like to happen
  - Can test for existence, probably can't measure it – does that help?
    - "I have a policy not to document everything". It's valid!
- Procedures: what you think should happen
  - Can test for existence, probably can't measure it – does that help?
- Practises: what actually happens
  - Can measure this,
  - But only at a given point in time
- Existence of policies, procedures does not mean they are followed
  - Who can guarantee the existence of an institution?

If we have partial compliance…

- We have some “level of compliance”
  - 1-gold star to 5-gold star
- Can we prioritise compliance requirements?
  - “What do I need for my first gold star?”
- Can we be more compliant in some areas than others?
  - “really nice policies, shame about the technology”
  - A single number can hide too much

Methods

- Who watches the watchers?
  - i.e. who measures the auditors?
    - Different auditors need to provide same answers given same inputs: Calibration
      - How much of the audit could we automate?
- Who keeps an eye on compliance?
  - Most elements involve humans
    - Compliance can be attained and lost, repeatedly
      - Maintain, review, test, and re-audit; trigger on changes to the audit report package?

Where does it stop, horizontally?

- Associated repositories for data movement
  - Federated repositories
- Data moves for
  - Performance (caching)
  - Protection (mirroring)
  - Policy (de-identification)
- *Outside* of my administrative domain
  - But strongly linked with it
  - How do I build trust in copies from authoritative sources? Does the local repository inherit some trust? Can a federation be made trustable?

Where does it stop, vertically?

- Designated Communities, domains
  - Want to trust the data
  - Need to trust the processes that created it
    - Which may be way before the SIP is built
    - 1-star lodgement effort into a 5-star repository? Or 5 into 1?
- Repositories can’t expect to
  - have sufficient domain expertise in-house, for evermore
  - be able to engage with a domain for evermore
    - Some domains didn’t exist before, or still exist!
  - deal with every format, software that a domain can use?
    - Unless you treat some of it opaquely?
  - Some of this should not be the repository’s problem

Where does it stop ??

- Problems with authentication, authorisation
  - External identity providers for authentication,
  - External policy providers for authorisation
  - How do we measure trust in them?
    - C3.3 has downstream obligation, but no upstream obligation?
    - Who takes responsibility that
      - policy is correctly expressed,
      - identifiers are correctly provided and
      - these things are correctly implemented
  - Documentation of accesses, modifications, using identifiers that may not be unique long term
    - re-use of usernames

Don’t we need positives *and* negatives?

- B5.2 has
  - *Review inappropriate “access denials”*
- But probably also need
  - *Review inappropriate “access approvals”!*

- C3.2 has
  - *Record accesses that “meet the requirements”*
- But probably also need
  - *Record accesses that “don’t meet the requirements”!*

Do we have 3 states of being?

- *Not compliant*
  - How could you be that bad??
- *Fully Compliant*
  - How could you be that good??
- *Partially compliant*
- Sufficient, in some/many cases?
- Users probably care about “just how compliant”
  - And depending on their relationship, different elements matter

What does partial compliance “mean”?

- i.e. who cares and why?
- 4 key players
  - Consumers
  - Providers
  - Funders
  - Repository Providers

Consumers care

- They want to trust the data
  - For each first-time access to a new dataset
  - For each recurring access to a particular dataset
  - Trust scope
    - the original data,
    - the process that got it in there,
    - the process that kept it there,
    - the process that got it out of there
  - Predictability, community, collaboration

- Probably only care about a fraction of the auditable elements, and care about some not-audited elements

Producers care

- Want the content to reflect what they provided
  - It's an additional cost to them to lodge data
- Want to leave a legacy
  - Collect once, re-use forever
- Want to gain recognition for the effort
  - Lodgement of scholarly input data as a form of publication
    - Requires a repository to be seen like a journal
- Probably care about most of the elements
  - May actually be a stronger relationship

Funders Care

- Need to trust the whole scholarly process
  - From research funding, through collection, to lodgement, and downstream re-use
- May be asked to recognise the effort
  - Or may enforce a requirement
- Requires measurement of value
  - Recognition is worth how much?
- Probably care mostly about how much the users care!

Repository Providers care

- What does it attract for them?
- Status as trust-able facility
  - To providers, consumers, and funders
- Supports arguments for ongoing support
  - How many repositories have guaranteed futures?

In closing

- I think this is crucial
  - Lots of things will be built on top of this
- I think this is hard
  - Lots of boundary issues
  - Lots of measurement issues
- I think this will all be solved
- I think this is all very very good