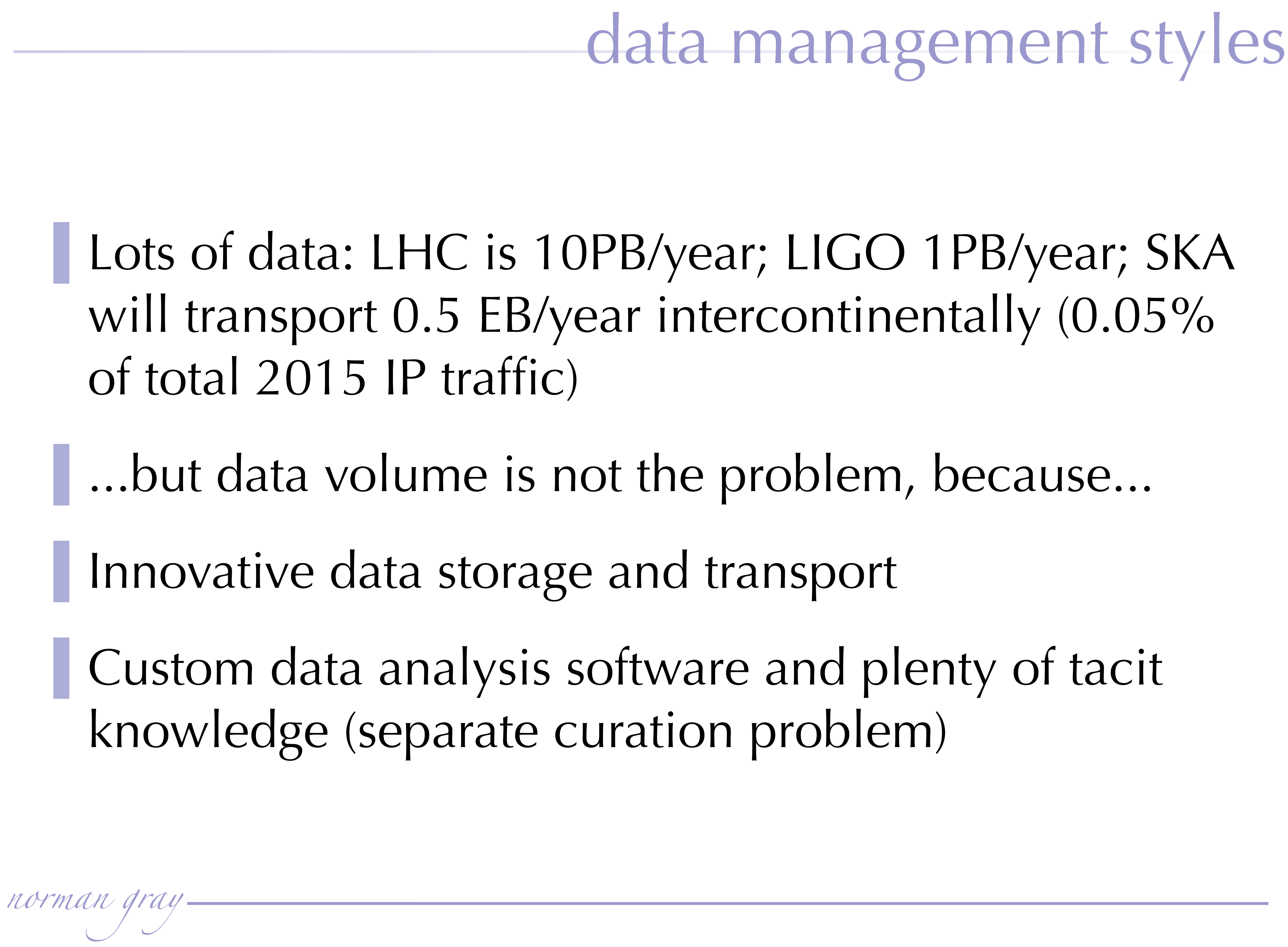
Supporting bespoke DMP in big science

Juan Bicarregui[1], **Norman Gray**[2], Rob Henderson[3],
Simon Lambert[1], Roger Jones[3], Brian Matthews[1]
[1]STFC, [2]Glasgow, [3]Lancaster

what is big science?

- big discoveries
- big money
- big author lists
- big admin
- big consensus

LHC+detectors approaching €10bn; LIGO ?US$1bn
ATLAS author list is 3000, LSC 600, so there exists consensus style management
Lots of data, but that’s not really a problem
MOUs, councils, workshops: engineering discipline – nothing ad hoc
If the community is persuaded something needs done, it’ll get done
Mostly talk about astronomy and HEP
HAVEN’T mentioned big data in this list



data management styles

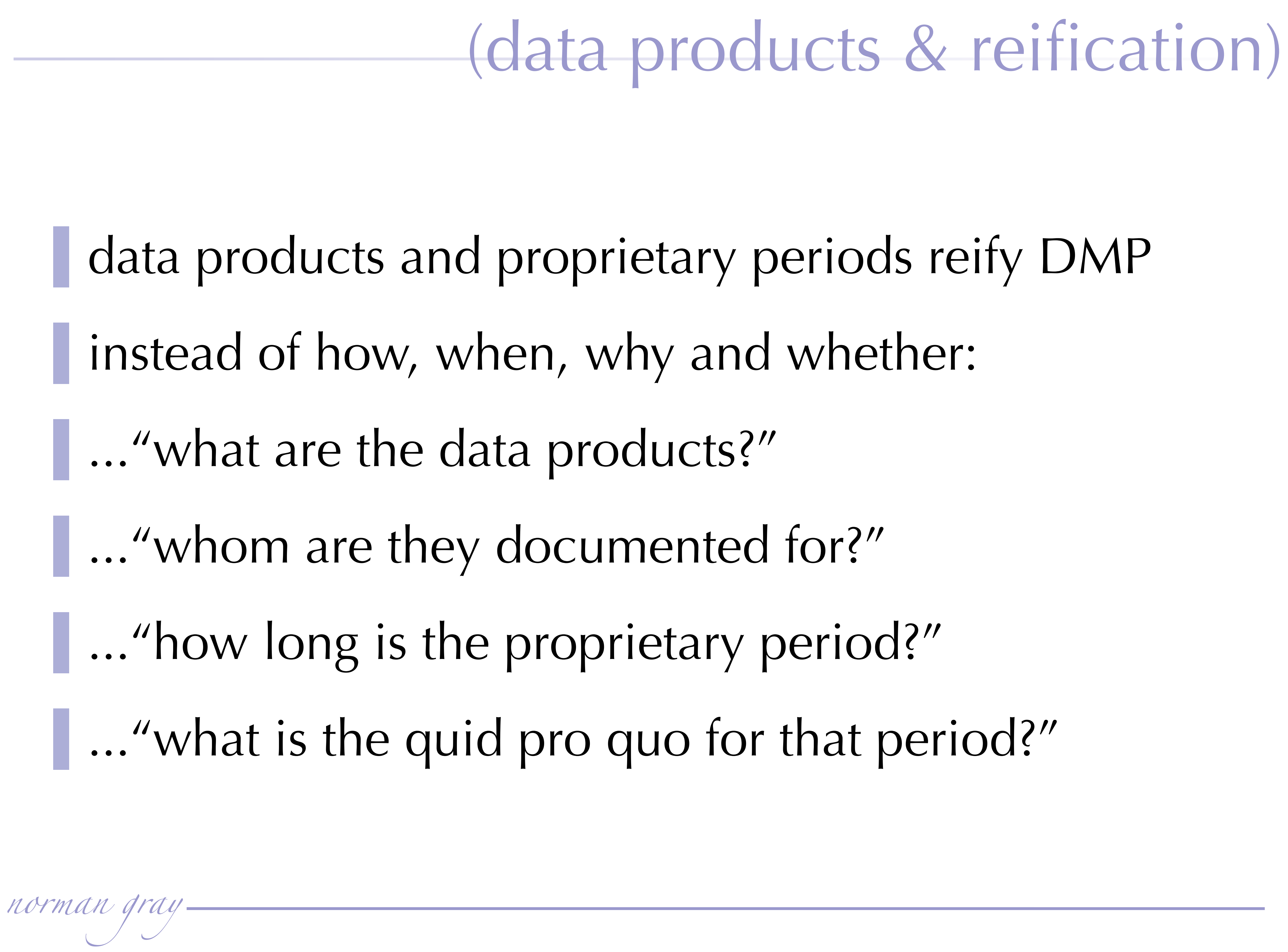
- Lots of data: LHC is 10PB/year; LIGO 1PB/year; SKA will transport 0.5 EB/year intercontinentally (0.05% of total 2015 IP traffic)
- ...but data volume is not the problem, because...
- Innovative data storage and transport
- Custom data analysis software and plenty of tacit knowledge (separate curation problem)

That is: this is strictly the left-hand side of the long tail

mrd-gw project

- The language of ‘data products’ and explicit ‘proprietary periods’ is useful
- Funders should simply require that a project develop a high-level DMP as a suitable profile of OAIS
- Funders should support projects in creating per-project OAIS profiles
- STFC should develop a costings model matched to the data challenges of the big-science community

Description of 'big science'
Intended to be of interest to JISC community, funders, and participants
MaRDI-Gross is support for 2 & 3, and a bit of 4



(data products & reification)

- data products and proprietary periods reify DMP
- instead of how, when, why and whether:
- …“what are the data products?”
- …“whom are they documented for?”
- …“how long is the proprietary period?”
- …“what is the quid pro quo for that period?”

Eg Herschel ‘science demonstration phase’: waive proprietary rights in exchange for more time

“Scientists should preserve and immediately share their raw data so other scientists, and the public, can reanalyse and reuse it”

well... up to a point, Lord Copper

Background to this is...
Encouraged by OECD, RCUK, STFC, EPSRC...
Complicated by international agreements, plus
...raw data is useless, and
...generating data products can be costly.
But good curation keeps these costs down
Echoing Jonathan, preservation-ready storage makes distribution easy/cheap

Summit of Mauna Kea, in Hawai`i. Left to right: Subaru (NAOJ), Keck I and II (Caltech), and the Infrared Telescope Facility (NASA)

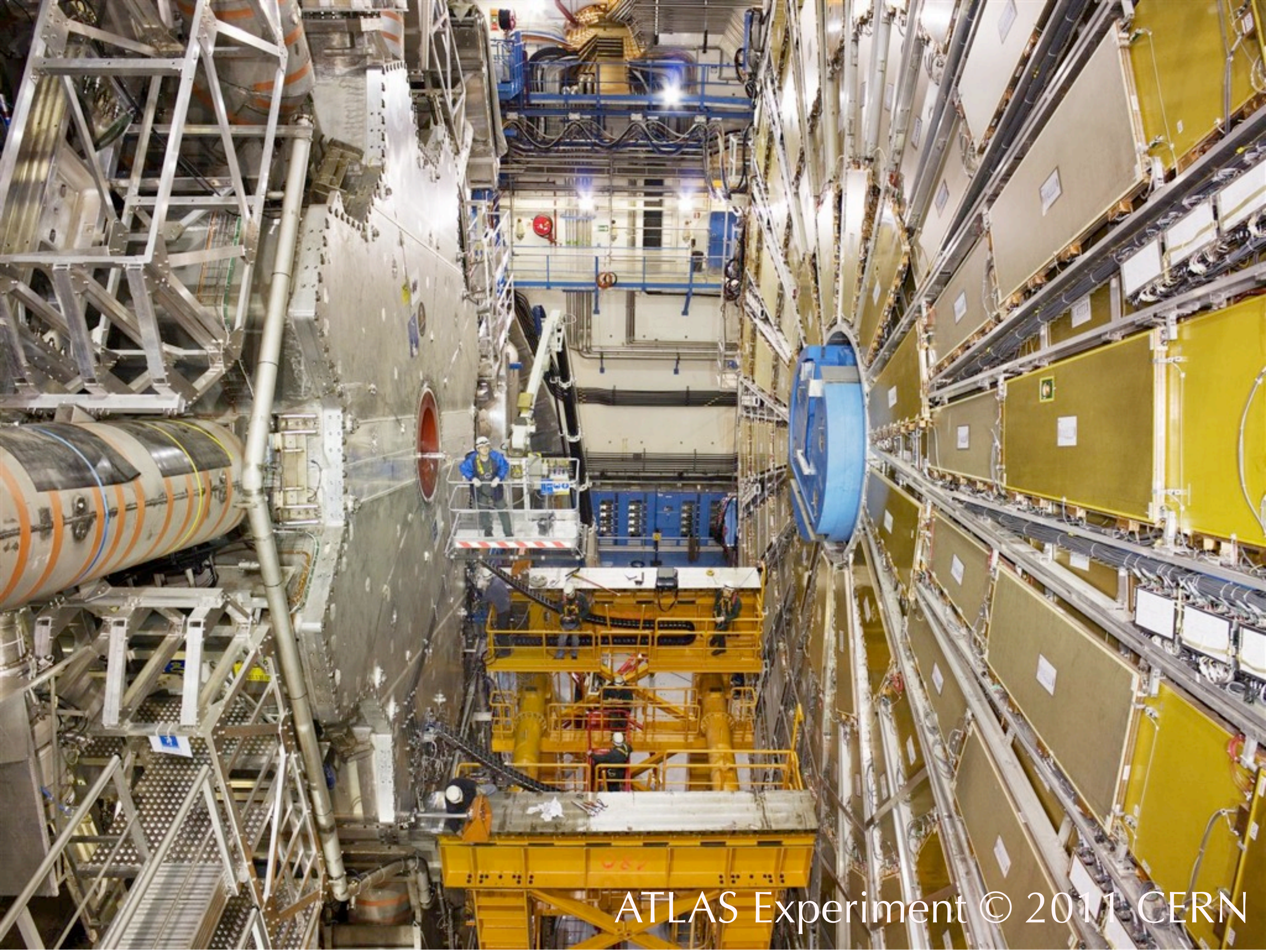
Mauna Kea: 13 telescopes; 11 countries --> International data ownership

Astronomy is an observational science (can re-observe plus transient events)

Largely intelligible data and common formats

Lots of experience of archive science

Data goes direct from instrument to archive --> Virtual Observatory



HEP organises measurements
So raw data hugely specific to instrument
...hugely specific to analysis chain
...and not shareable or easily preservable
Data preservation is not a no-brainer, so what’s reasonable?
multiple PB/sec decimated to ATLAS 10PB/yr

MaRDI-Gross

Managing Research Data Infrastructures – Big science

Building on MRD-GW recommendations

Case studies/experience in ISIS (pulsed spallation neutron source); multiple instruments and multinational users

Gravitational waves: 1300-person consortium in US and Europe, currently in 2011-15 upgrade, PB/year when running, currently has an OAIS-style DMP plan

ATLAS and the other LHC experiments: 10PB/year; very actively discussing data release planning with STFC and others

Introduction Focuses, coverage, and some definitions • The what, why and how of OAIS • What is 'big science'? • What is 'data management and preservation'?

Policy – the 'why' of DMP planning RCUK data principles and their interpretation • Sharing: openness and citation • The argument for open data • The argument for data preservation • Should everything be preserved?

Technical background OAIS • Preservation Analysis in CASPAR • Audit and certification of trustworthy digital repositories • The DCC curation lifecycle model – a contrast to OAIS

DMP planning – practicalities Preservation goals • Data release planning • Validation • Software and service preservation • Costs and cost models • Modeling storage costs • Modeling data loss

Case studies in preservation ISIS • LIGO/GEO/Gravitational Waves • LHC experiments

STFC Data principles

Table of contents
Policy: projects may want/need to push back to funders
Technicalities: underlying technologies rather than implementations
Practicalities: miscellaneous practical details

claim

- The demand for principled data management and data sharing is a reasonable and shared one;
- a reasonable framework for at least approaching the problem already exists in OAIS;
- the OAIS recommendation is (just) concrete enough that it is not merely waffle; and
- there is a bounded set of resources which will allow DMP planners to produce a practical project DMP plan, reasonably painlessly.

So our idea of funder DMP guidance is ...

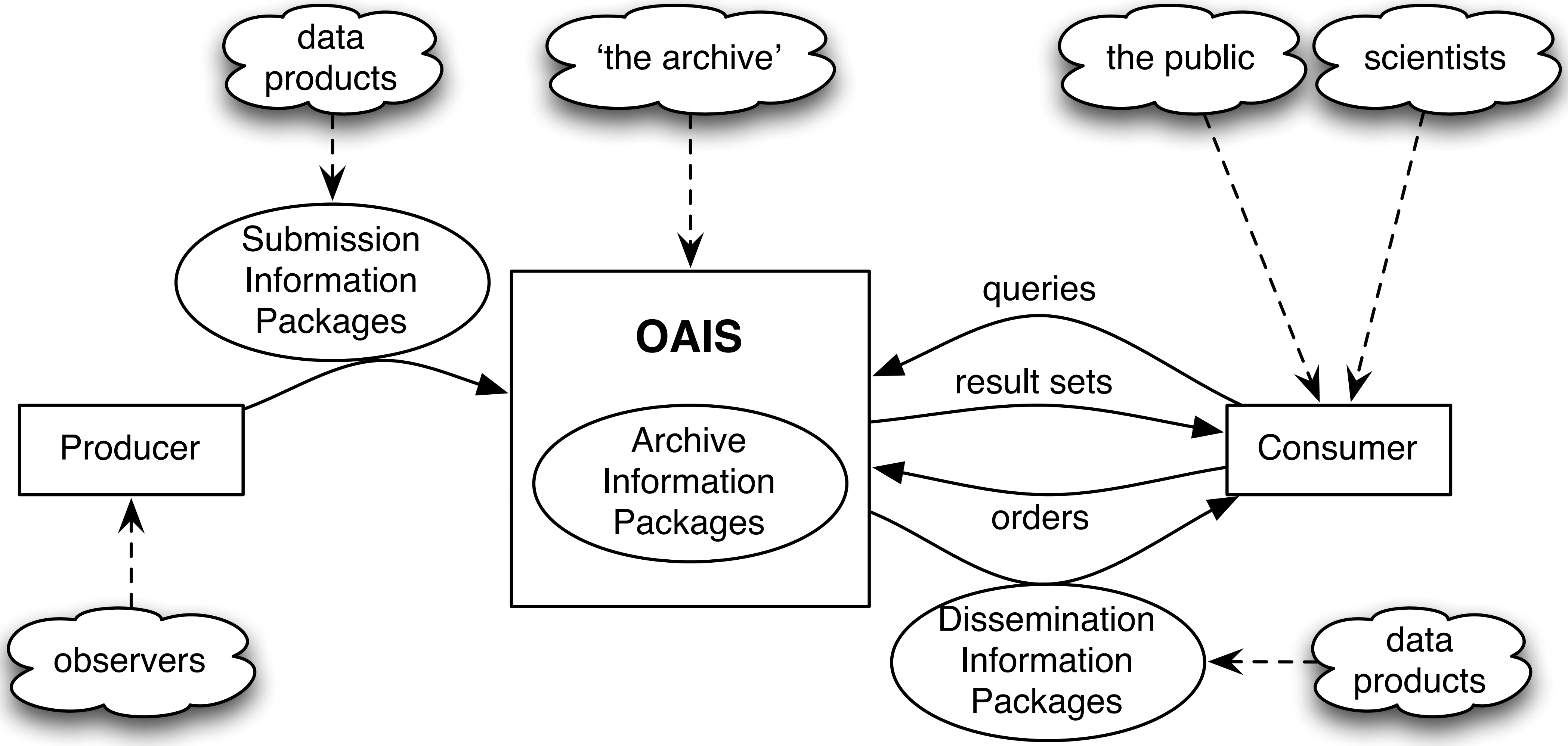
Here’s a copy of CCSDS 650.0;

it’s sane;

get on with it

...as the funder should say to the project

OAIS



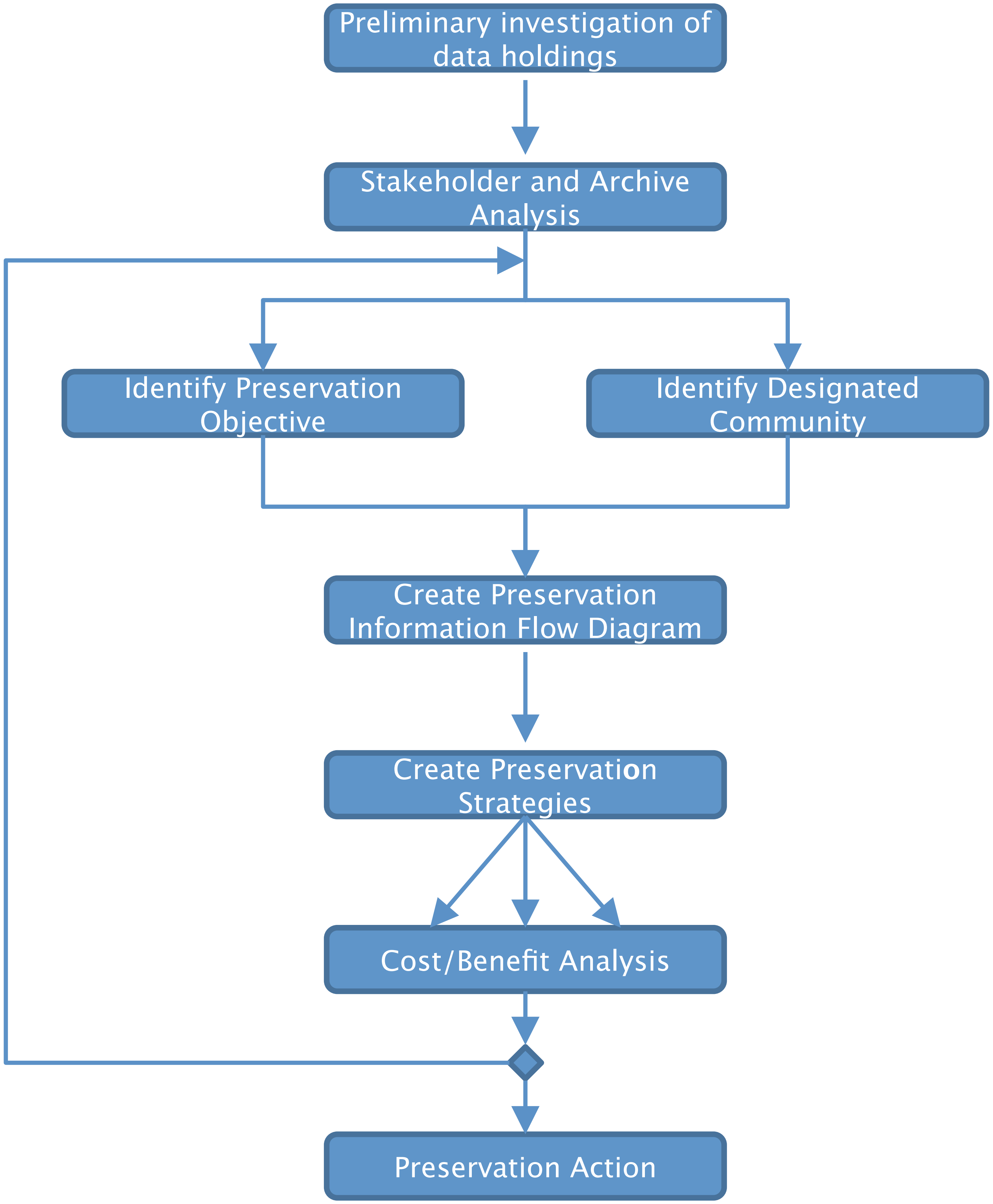
CCSDS 650.0 (2002) = ISO 14721:2003

* a Standard – more importantly, a common language supports conversation
* about the flow of data into, within and out of an archive
* facilitates conversation with funders
* describes/formalises what's common practice in space data (therefore realistic)
* set of concepts and terms (so interop)

“almost any system capable of storing and retrieving data can make a plausible case that it satisfies the OAIS conformance requirements”

Rosenthal et al
Saying “we promise not to lose this USB stick” can probably be dressed up in OAIS party-clothes, but isn’t really a plausible DMP strategy

CASPAR


Preliminary investigation of data holdings
Stakeholder and Archive Analysis
Identify Preservation Objective
Identify Designated Community
Create Preservation Information Flow Diagram
Create Preservation Strategies
Cost/Benefit Analysis
Preservation Action

CASPAR
Methods and tools for several stages of the DMP lifecycle
Examined cultural heritage, performing arts, and science data, for validation
The point is to create a structured process
...plus software tools

‘Trustworthy repositories audit and certification (TRAC)’

↓

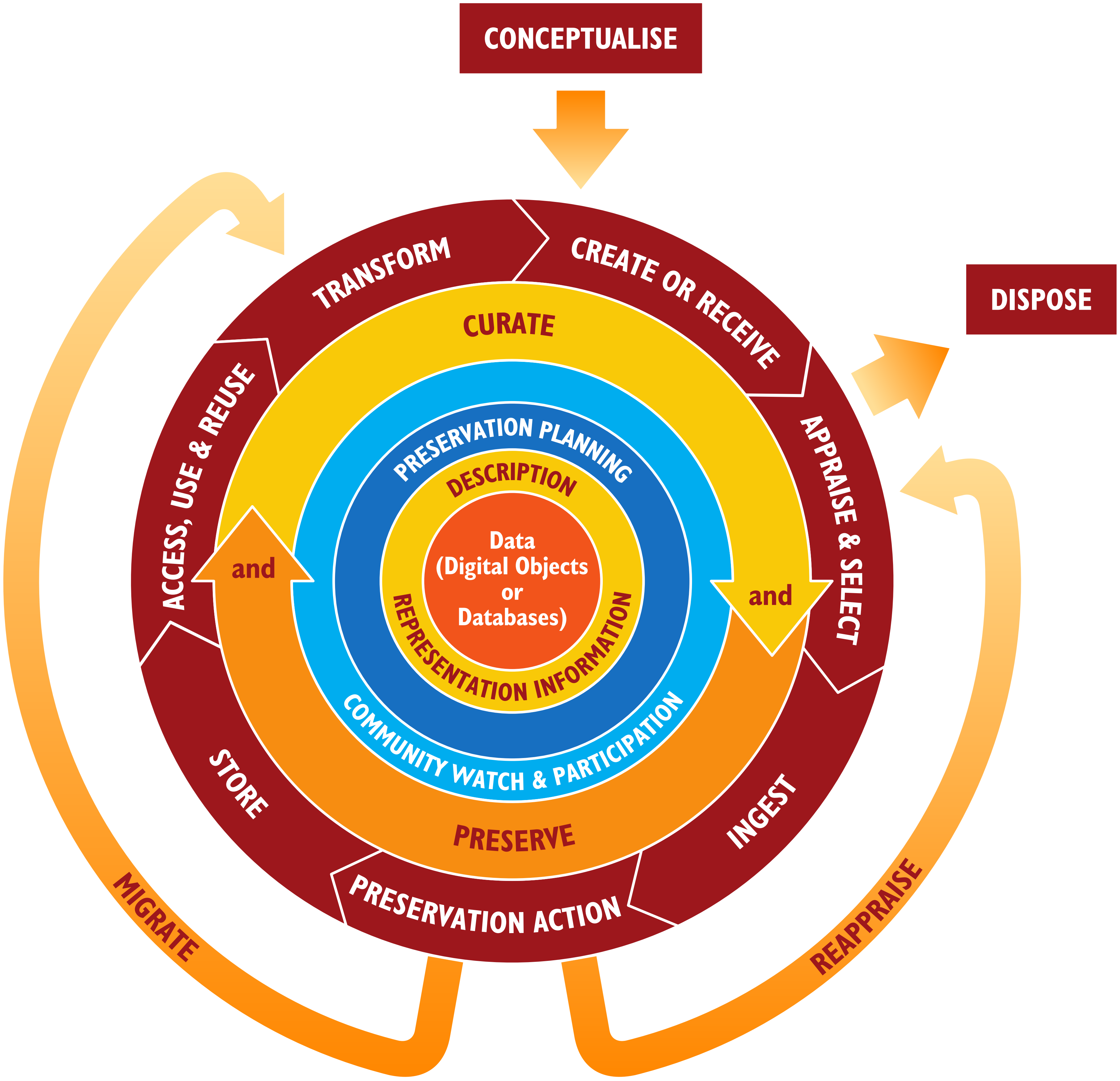
‘Audit and certification of trustworthy digital repositories’
CCSDS 652.0 = ISO-16363:2012

Audit and certification
A number of repository metrics: 'The repository shall...' + examples + rationale

validation

- TRAC is considering ‘bronze’, ‘silver’ and ‘gold’ certification
- Should funders require, say, ‘bronze’ certification for projects above a certain scale, or in certain circumstances (eg climate data)?
- This will probably not be hard for a well-run big-data project. Self-certification might be adequate
- ...but not free.

Big-data projects need good data management in order to run successfully
(Also encourages projects to engage with TRAC, and perhaps this should be encouraged/funded by RCs)



OAIS isn't the only show: DCC has an alternative model
"very repository-centric" (don't show this to data owners!)
(That is, it's more important that a DMP planner has SOME model, than that we obsess about what that model is)

Practicalities

- Intended to not be just waffle
- preservation goals: who's in the DC? What will they want to do? For how long?
- No generic answers
- and 'how long' makes a big difference, since it can potentially underestimate both the costs and benefits of long-term preservation

release algorithms

- LIGO: explicit algorithm for timing data release, which is a function of time, amount of space explored, and discoveries
- ATLAS has ‘RECAST’ service: they don’t release data, but will re-analyse their data with your model
- Astronomy: either proprietary periods, or surveys have periodic DRs after QA checks

- Some experience of plausible round numbers in the area; hard to make robust estimates until well into the project
- Storage: scarily expensive
- Rosenthal models, based on Kryder's law, which describes the continuing decrease in the cost/MB of storage
- but is it decreasing quickly enough?
- endowment of $3000/TB

- ingest: often very expensive, and front-loaded; can be messy; generally absorbed in infrastructure costs for big science
- staffing: expensive but predictable

purl.org/nxg/projects/mrd-gw

purl.org/nxg/projects/mardi-gross

nxg.me.uk

MRD-GW report still live
MaRDI-Gross finalised at end of year