Organisational provenance capacity implementation plan

At Geoscience Australia

Peter Fitch, Nicholas J Car & Sue Fyfe

What I want to talk about

- Why Geoscience Australia is interested in provenance?
- How this work came about and what the key questions were?
- A look at the implementation plan developed
- Next steps

Lessons of Climate Gate

Theft of e-mails from UEA Nov 2009

E-mails indicated manipulation of data, and suppression of raw data

Investigations found

- methods dis-organised
- bunker mentality
- lack of transparency

Researchers promised to

- improve scientific data management
- open access to data
- Improve transparency

Geoscience Australia

Science Principles

Contents

Provenance in QA/QC

Quality Assurance (QA)

Quality Control (QC)

Principles
Policies
Procedures

"What to do and why"

Data product specification

Provenance

"What Was done"

Objective of the project

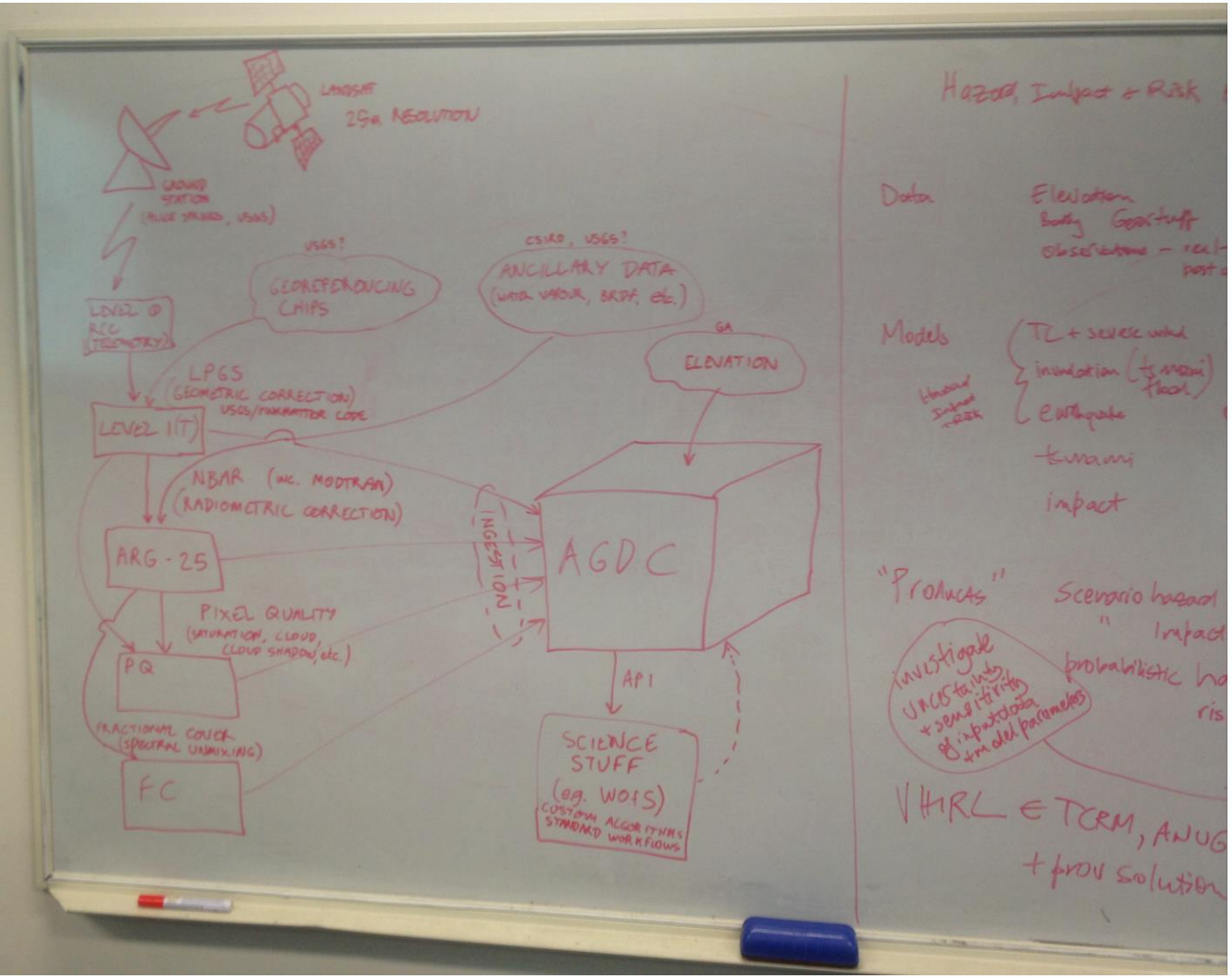
Answer 2 questions:

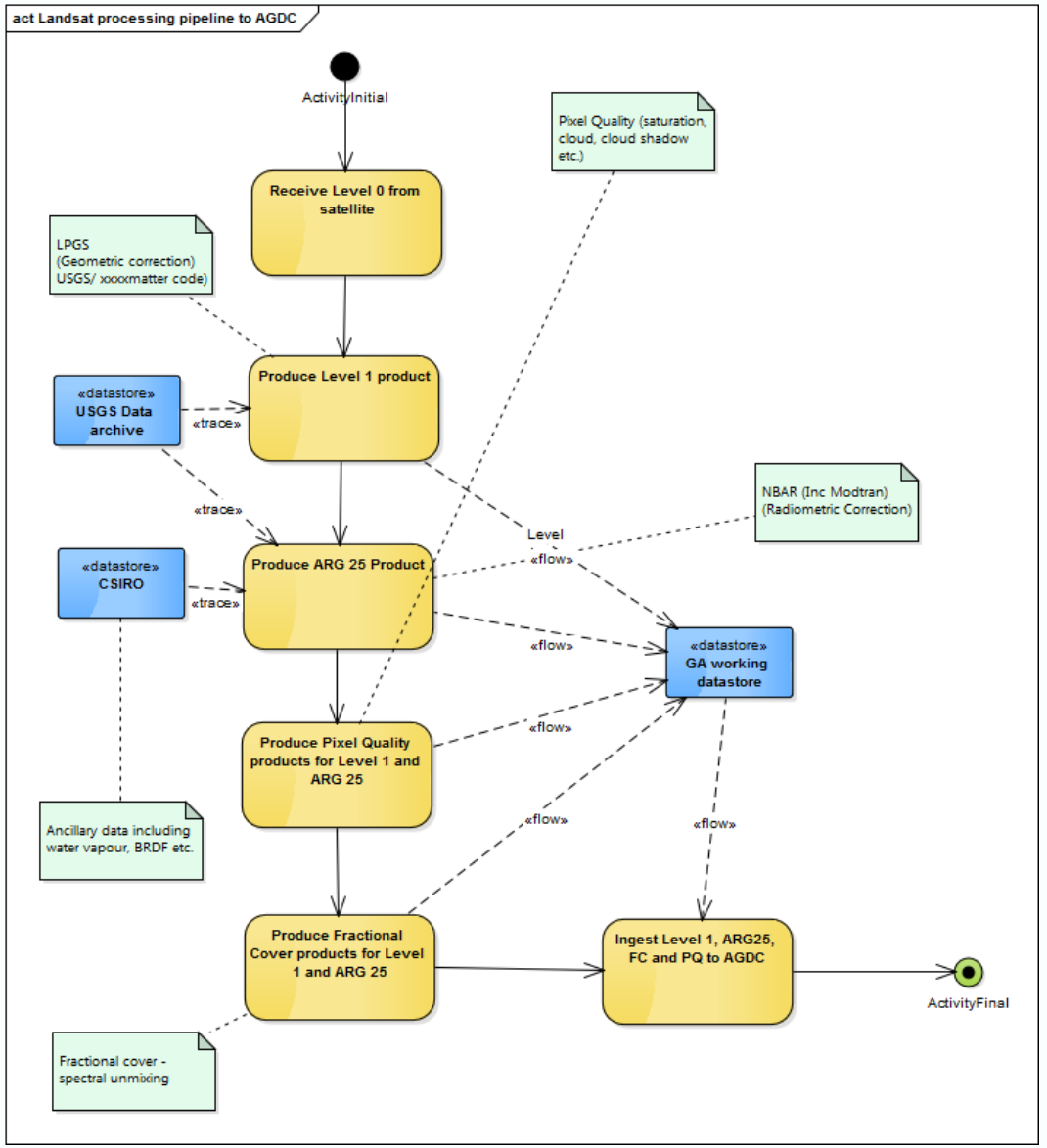
- To confirm that a consistent approach to the capture and use of provenance information across GA is possible and sensible and,
- If so, provide GA with a recommended work-plan with sufficient detail suitable for progressive implementation.

This required:

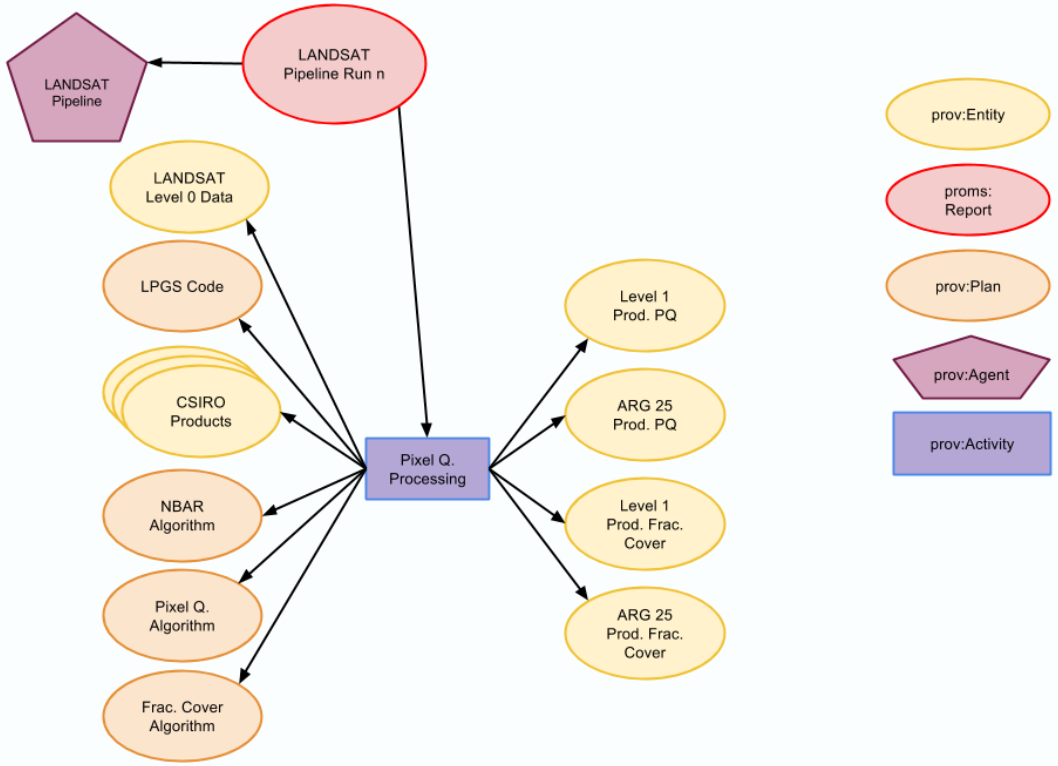
- Developing an understanding of the range of workflows undertaken across the business;
- Confirm (or not) that a standard approach to dealing with the capture and use of provenance information across the range of workflow can be achieved, by analysis of the workflows;
- Identify the social/institutional barriers to implementation, if any, with recommended actions to overcome them;
- Developing a technical architectural solution.

Understanding the workflows

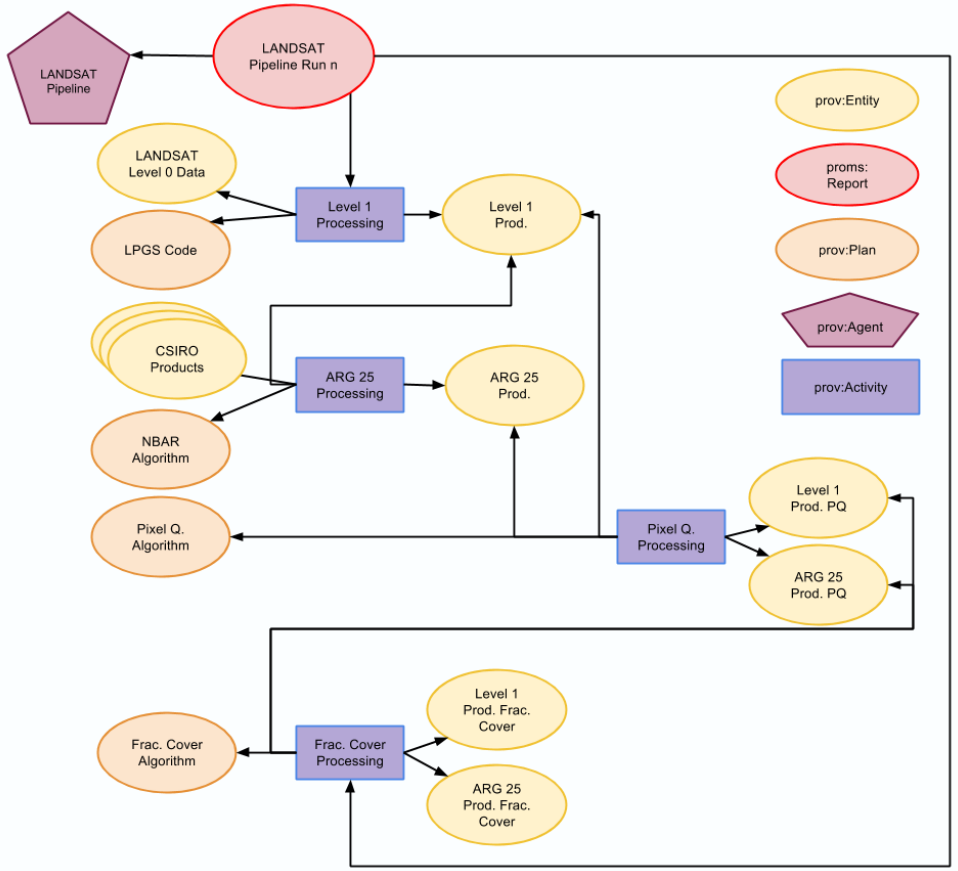




Mapping workflows to Prov-O



External view



Internal view

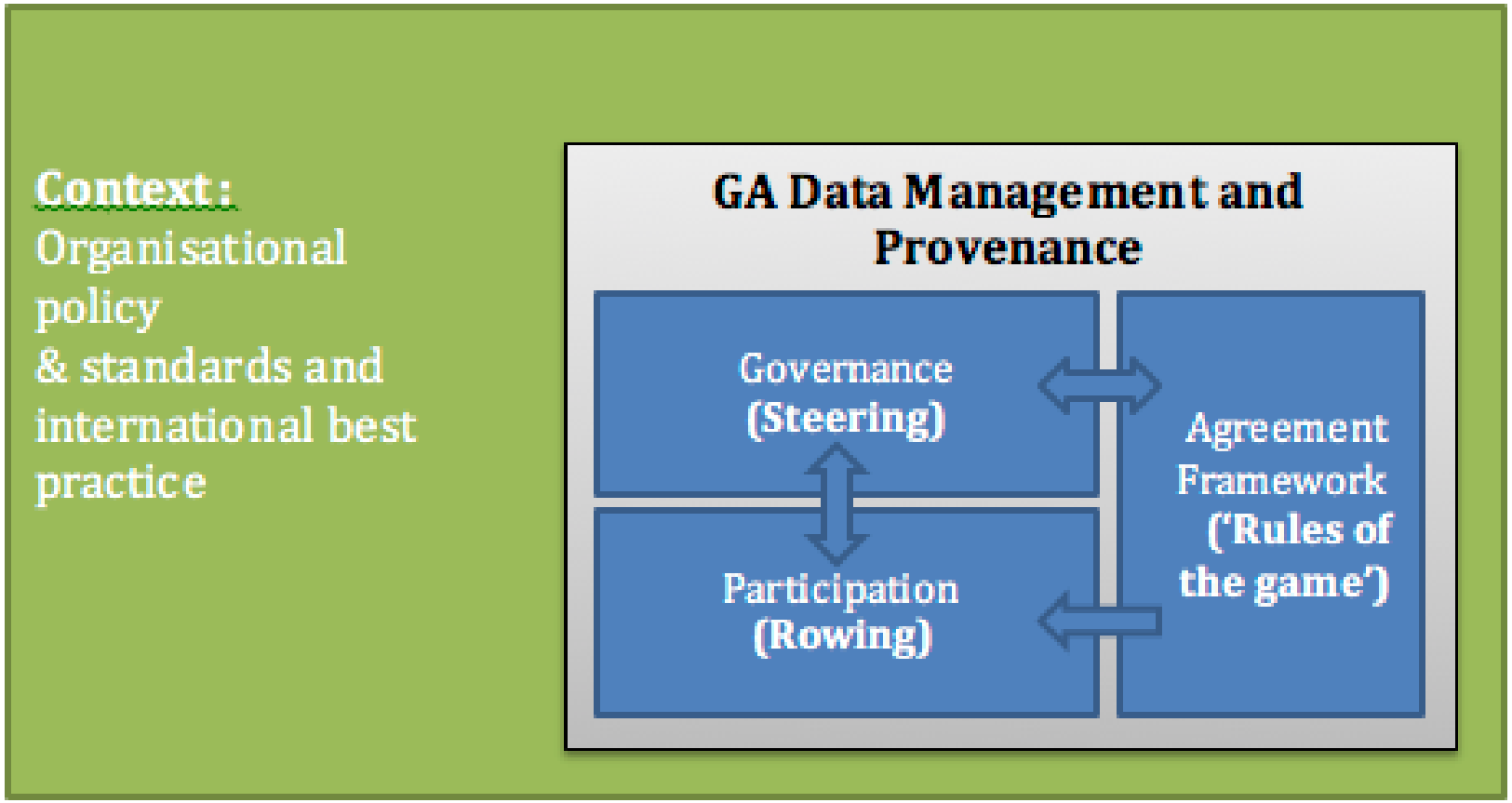
Can GA workflows be mapped to Prov-O?

Workflow Name	Description	Categorisation (C, HC, H)	Mapped PROV-O?
Landsat satellite data processing pipeline	This is the automated processing pipeline for the Landsat data stream	C	Yes
EMA exposure report	A human-computer workflow which is run daily to provide assessments of financial and social exposure to natural hazards including fire and flood.	HC	Yes
Sentinel hotspots data	This is a fully automated workflow that generates hotspot datasets and are derived from satellite observations that detect heat on the land surface.	C	Yes
Provision of expert advice	A fully human workflow, where a GA expert is contacted by government or industry to provide expert advice. This advice is typically presented in the form of a letter or report.	H	Yes
Laboratory workflow	An extremely varied mostly human workflow analysing samples for various chemical and physical properties.	H	Yes

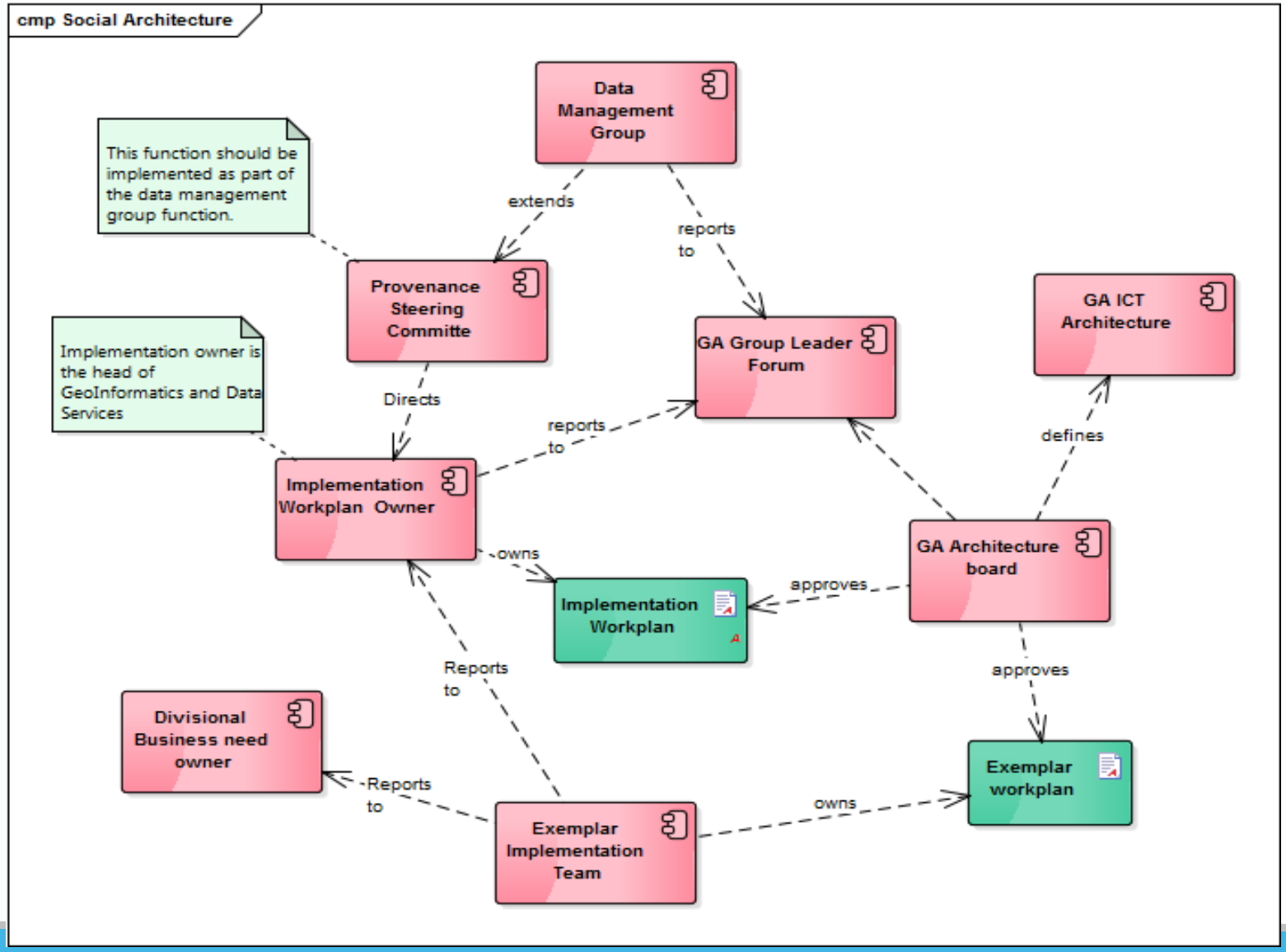
- C – Computer
- H – Human
- HC – Human Computer

Yes – Prov-O can be use for GA workflows.

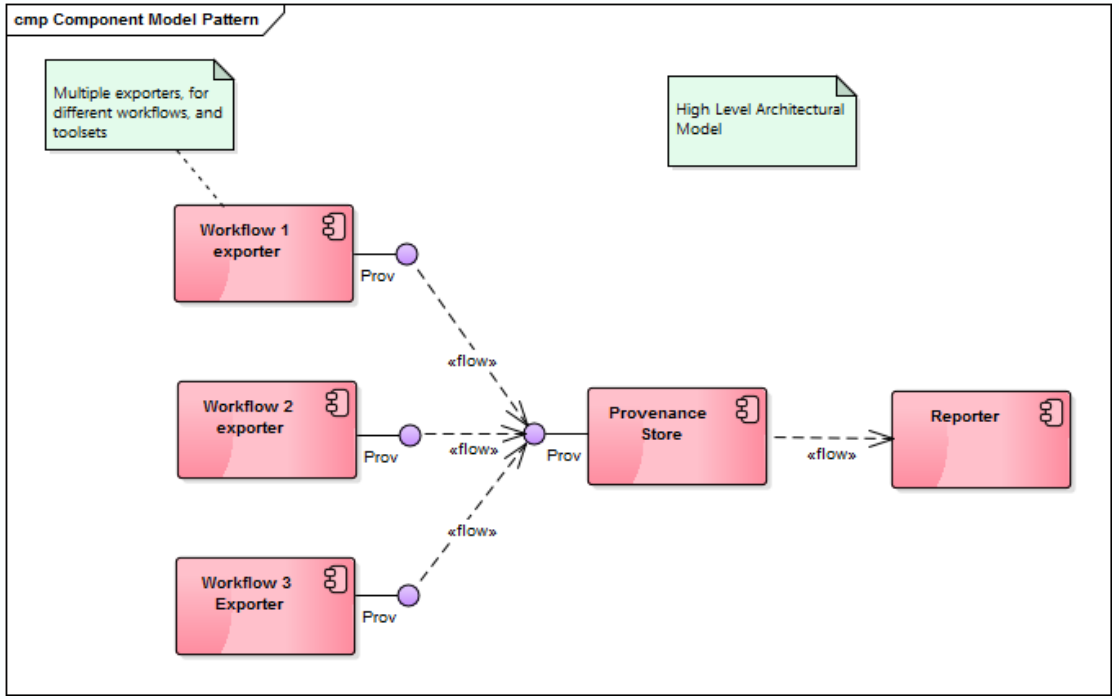
The socio-technical problem



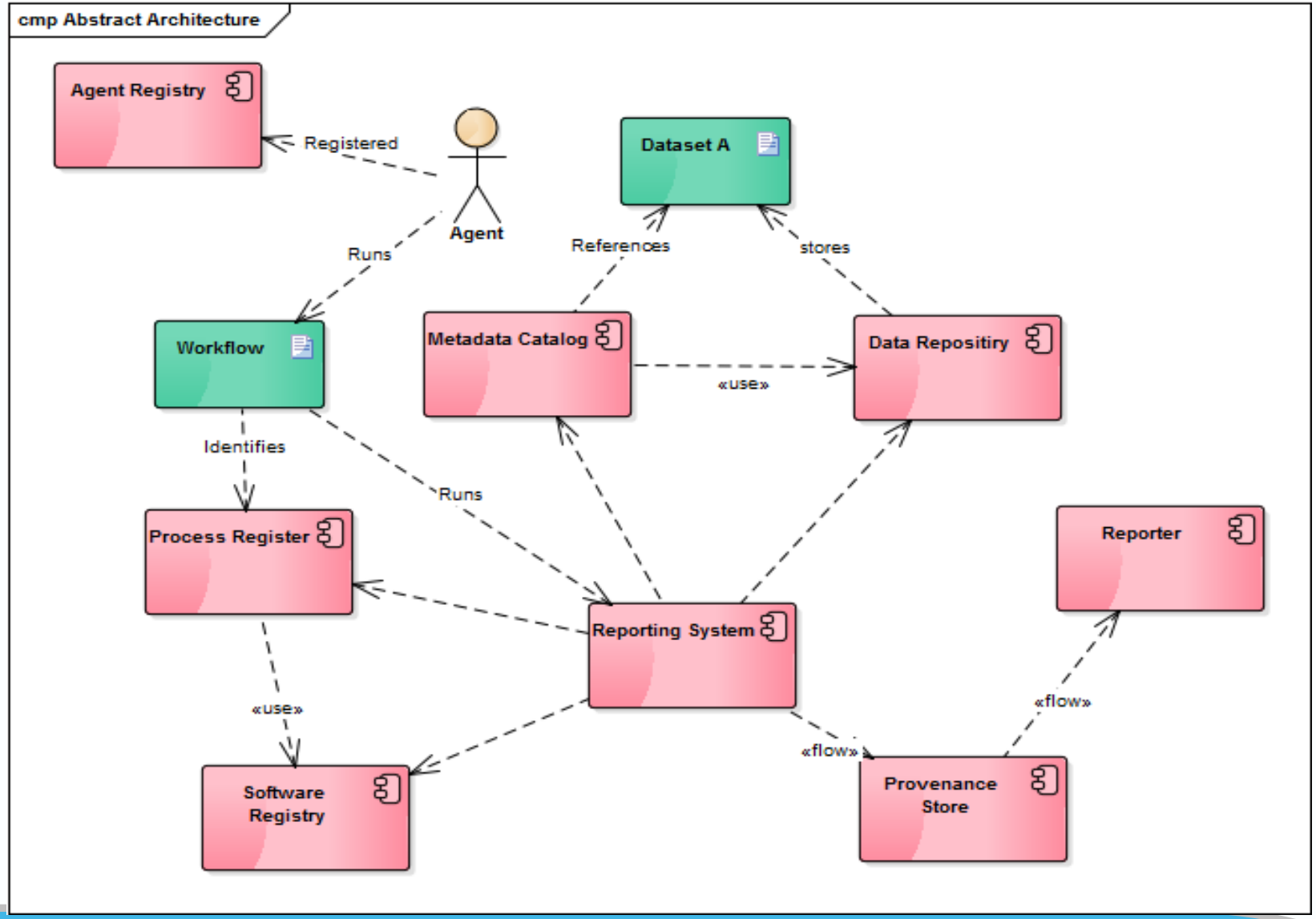
A social architecture



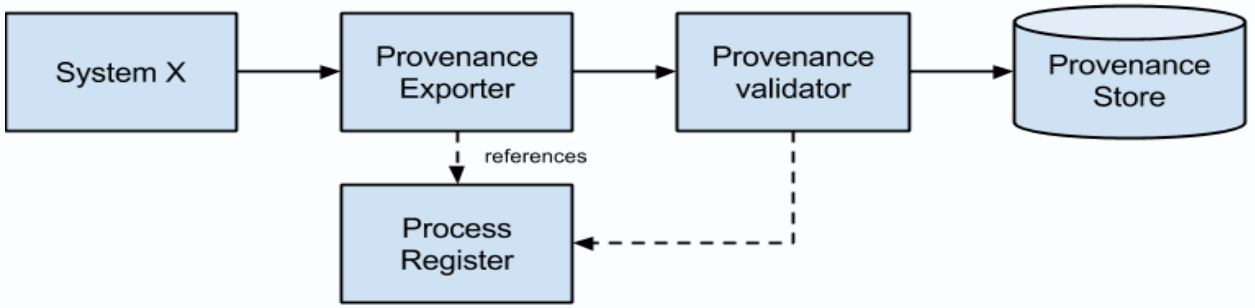
Abstract Architecture



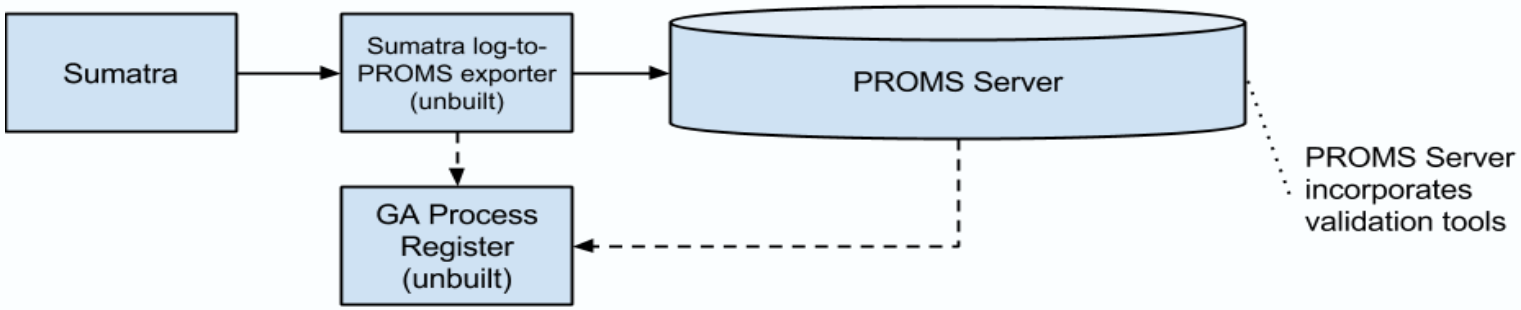
System Architecture



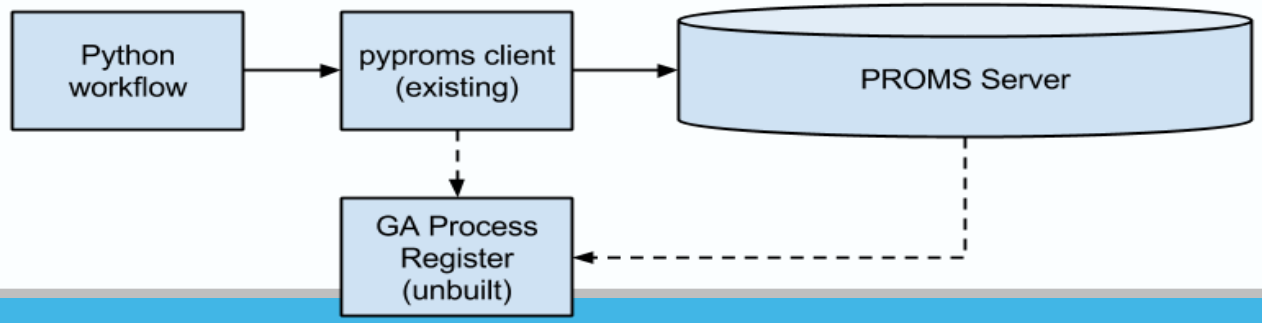
Platform-independent model



AGDC platform example #1 - Sumatra



AGDC platform example #2 - Custom Python workflow



Identity Backbone.

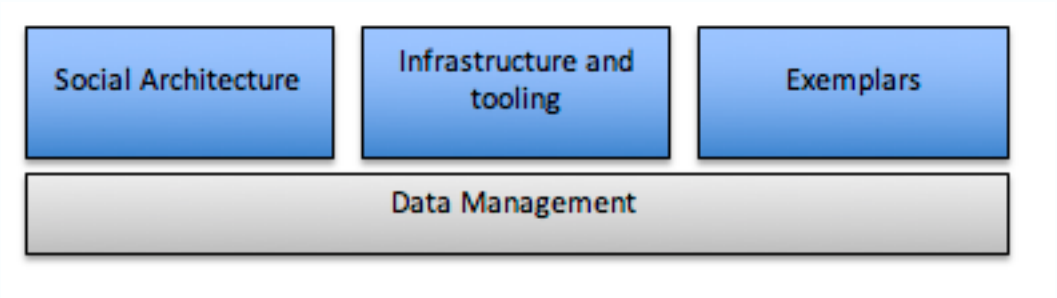
Thing	Description	Identity	Registry	PROV-O mapping
Dataset	A digital data object identified by a URL	URL	Metadata Catalog	prov:Entity
Software	A set of software codes identified by a URL	URL	Process	prov:Entity or prov:Activity
Script	A script used to embody a workflow identified by a URL	URL	Process	prov:Activity or prov:Entity to prov:Agent
Model codes	as for software	URL	Process	as for software
Algorithm	as for software	URL	Process	as for software
Workflow	a workflow which operates on data and creates data identified by a URL	URL	Process	prov:Activity or prov:Entity or prov:Agent
Person	an actor who is responsible for the execution of a workflow identified by a URL	URL	Operator	prov:Agent
Web service data use	Data taken from a GA or other organisation's web service and used as per dataset use. Identified by a URI and metadata.	URI + metadata	Potentially a Web Services Register (if a GA web service) or no register - represented only as an element in provenance traces	prov:Entity at a basic level with notes but better as a sub-classification proms:ServiceEntity

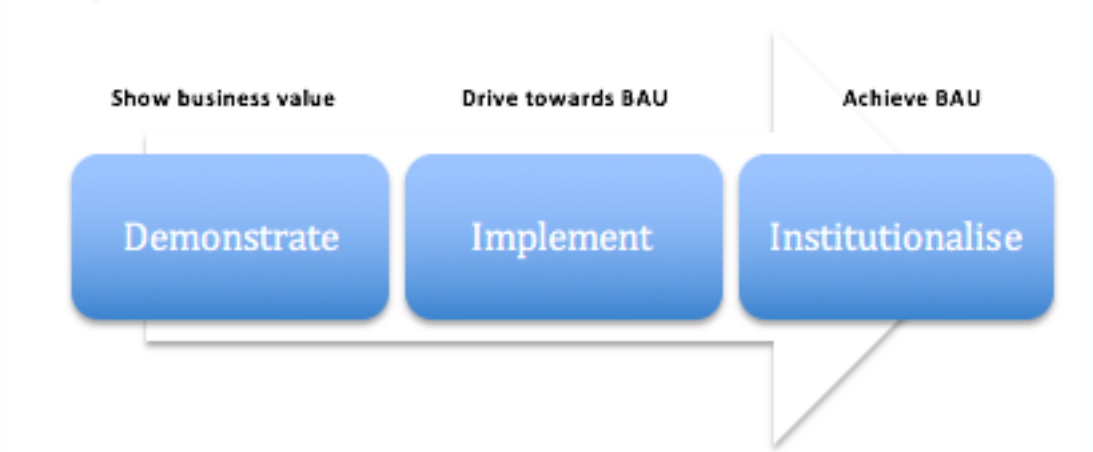
Plan Overview

Year 1 - To demonstrate - build basic infrastructure, undertake exemplars, and review progress;

Year 2 - To Implement - extend to more exemplars and harden infrastructure, extend toolkits and review progress;

Year 3 - To Institutionalise - further add exemplars and review progress.





KPI's – How do we know we're improving?

- **Provenance Maturity Model score** – PMM evaluation.
- **Product provenance Audit results** – Audit products for provenance
- **Total number of datasets** - with and without metadata;
- **Total number of provenance records** - in the provenance store;
- **Total number of Activities registered -** in the Activities register;

Data usage statistics -looking at the usage of data with metadata and provenance versus data without;

List of the projects that require provenance information as part of the contract.

Number of documented scientific processes, algorithms, software: that are accessible and available online for reuse e.g. in VLs;

Number of exemplar projects complete.

Parting Thoughts - conclusions

1. We have developed a provenance information capture and use implementation plan.

2. It tackles both the technical and social aspects of implementing a provenance system.

3. We found that Prov-O can cope with the range of workflows found so far with geoscience Australia.

4. Abstract architectural patterns provide a back bone to implementation.

5. Come back to MODSOM in 2017 to see how good the plan has been!!

Thank you – Happy for questions

CSIRO Land & Water
Peter Fitch
Senior Experimental Scientist
peter.fitch@csiro.au

Geoscience Australia
Nicholas Car
Data Architect
nicholas.car@ga.gov.au