Department of Natural Resources, Mines and Energy

Geological Survey of Queensland

“Data for Discovery”

Unlocking the full value of geoscience data to enable industry success

Queensland Government

BIG Data Revolution - Global Shift

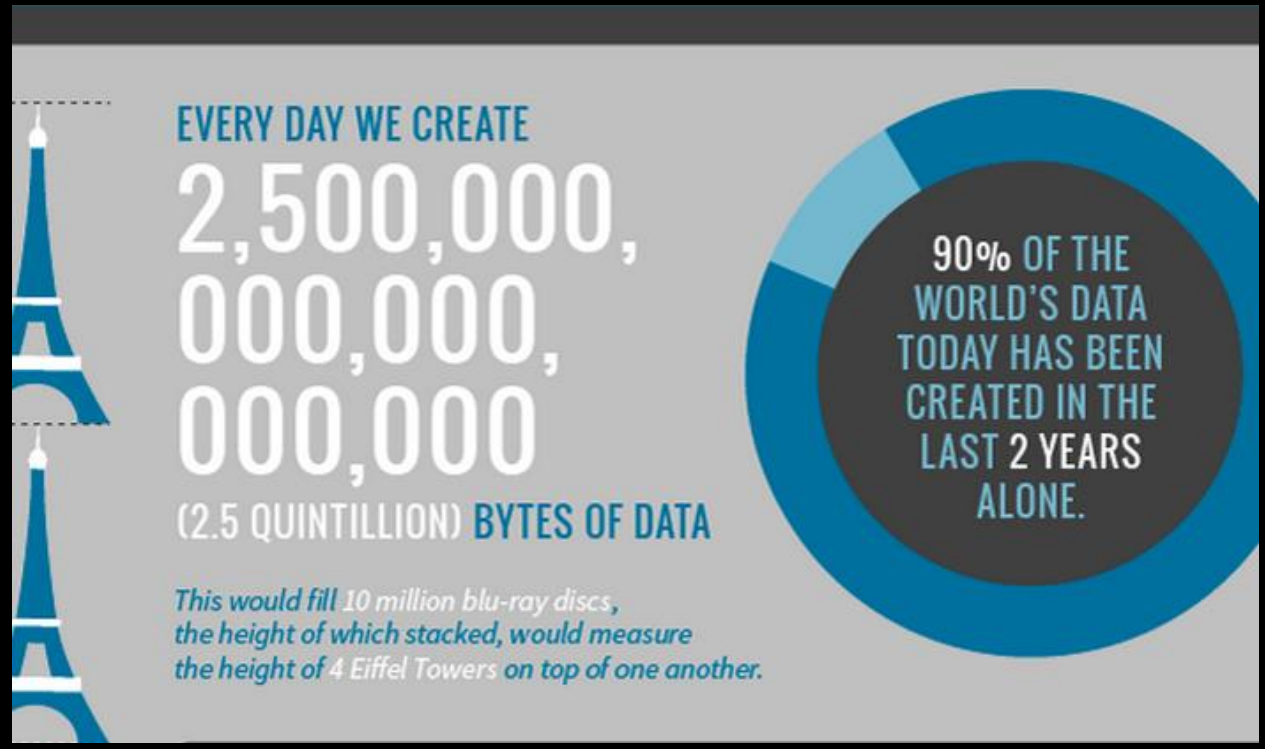
- Big Data Revolution – fuelling a global change
- We are at the beginning of a new information age
  - Driven by;
    - Cheap internet connectable sensor technology
    - Massive connectivity jump – internet expansion 2005 (1B users) – 2018 (4B users)
    - Organisational culture change to more open sharing of data and information

Resulted in access to massive amounts of data –

Big Data!

Big Data Revolution – Some Stats

- 90% of the world's data has been created in the last two years alone.
- Most companies only analyse 12% of the data they have.
- By 2020, there will be more than 50 billion smart connected devices in the world, collecting, analysing and sharing data.

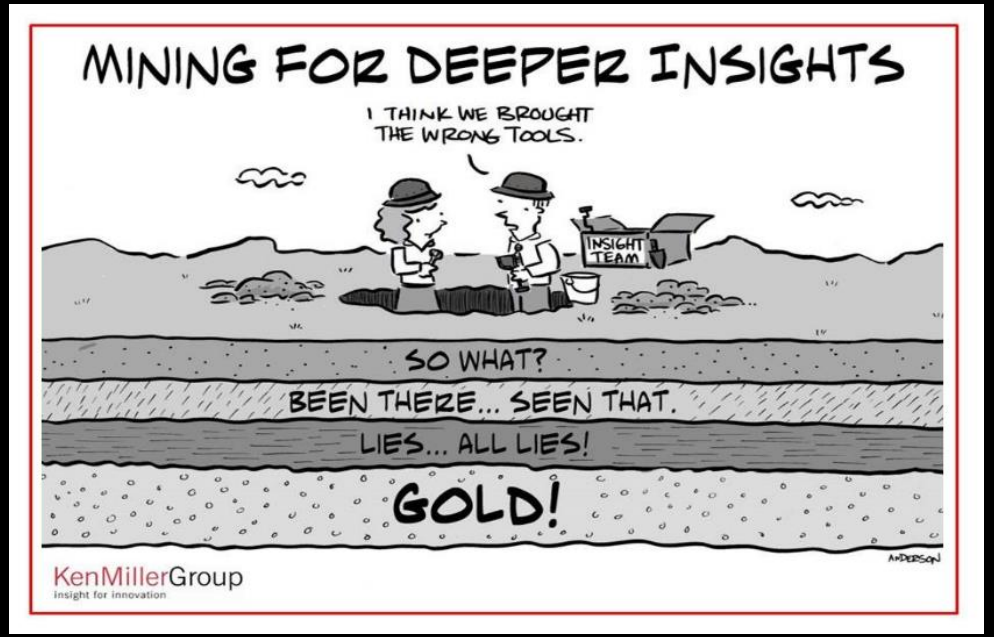


- Bad data costs the US $3.1 Trillion annually.
- IoT will save consumers and businesses $1 trillion a year by 2022.

https://www.attunity.com/blog/10-eye-opening-stats-about-the-growth-of-big-data/

Big Data – The Insight

- Not just about collecting data for data's sake
- Large scale data access supports the capacity for large scale analysis
- It's not all about data resolution
- It's also about finding relationships between different data sources…



Analytics + Data allows us to see <u>NEW</u> patterns…..

Why is this important?

- Access to data and analytical techniques is driving significant improvements in many other domains;

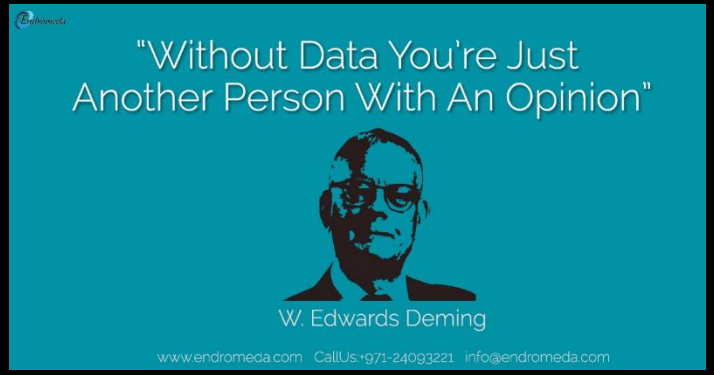
Breast cancer affects 117 out 1000 Queenslanders every year.

What if I told you we could improve detection rates a further 16% by simply using the same data more effectively...?

What if we could save 10% of Queensland Rail's $1.1B annual operating cost by just collecting and using data more effectively...?

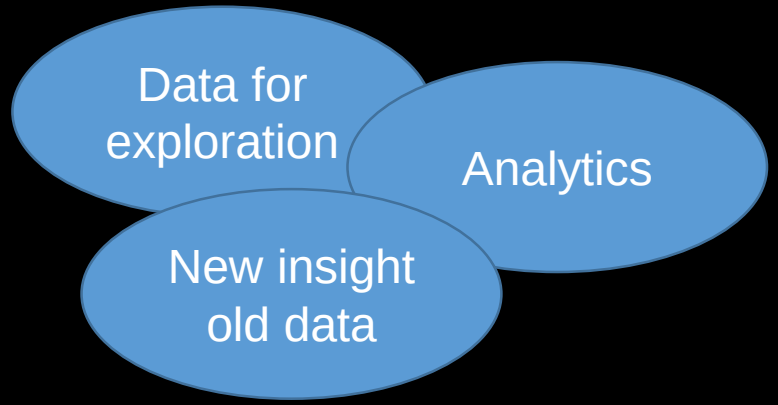
Some Examples

- Some of the massive opportunities in just using data more effectively…
  - Google Analytics has improved the breast cancer detection rate from 73% to 89% in the US - using biopsy data and machine learning
  - Norfolk Southern (an eastern US rail network provider) optimised it's network and it saved the company 200 million dollars per annum – using sensor data and machine learning.
- It’s a big deal in other domains too…



Geoscience Data Modernisation Project

- So what does this have to do with Geoscience?
  - The “Big Data” global revolution is driving change in the Geosciences just like it’s driving change in healthcare, logistics, insurance, advertising – the list goes on….
    - Even greater focus on the use of data in exploration
    - Leveraging big data analytical techniques for geoscience problems
    - New insight from old data
    - Data and insight underpinning decision making – better decisions
  - In turn this drives;
    - Change to improving accessibility, quality and reusability of our existing data and information
    - To fuel analytics programs more effectively



Early Adopters

- Already being embraced in the geoscience world…
  - Using data and analytics GoldSpot Inc. was able to find 86% of the existing gold deposits in the Quebec Abitibi, but only needed 4% of the total surface area to do so.
  - Earth AI – developed machine learning software which generates precise exploration targets in previously unexplored areas.
- Access to both novel data sources and analytical power gives us;
  - Gives us the ability to explore more challenging areas – especially when the potential target is undercover.

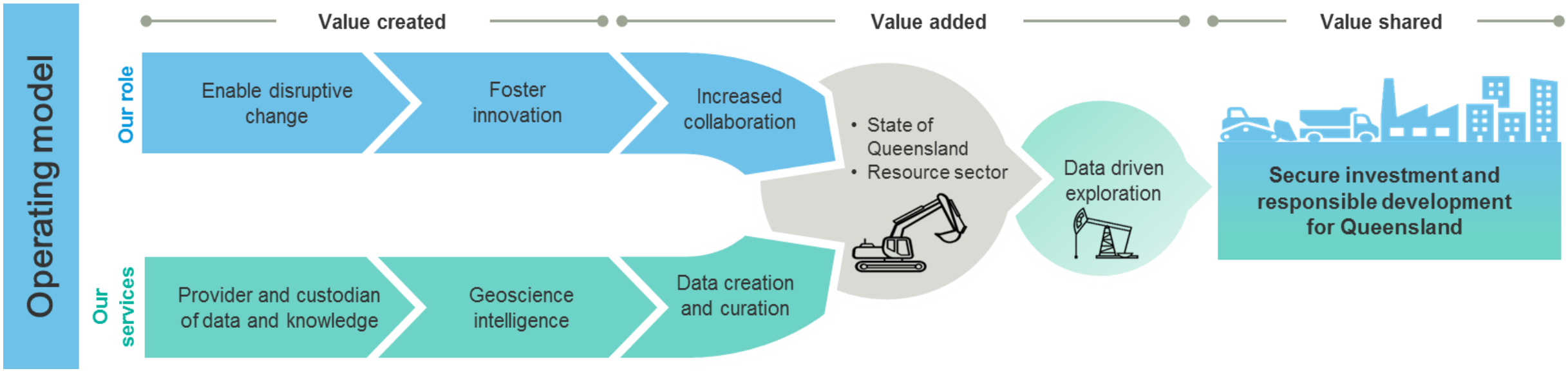
Success requires smarter ways in not only redefining how we use data, but how we manage and curate it to ensure it can be used for geoscience problem solving.

80% of the effort is in data management!

GDMP – Data for Discovery

GDMP will deliver a smarter way to manage and access GSQ custodial data.

GSQ's Strategic Direction



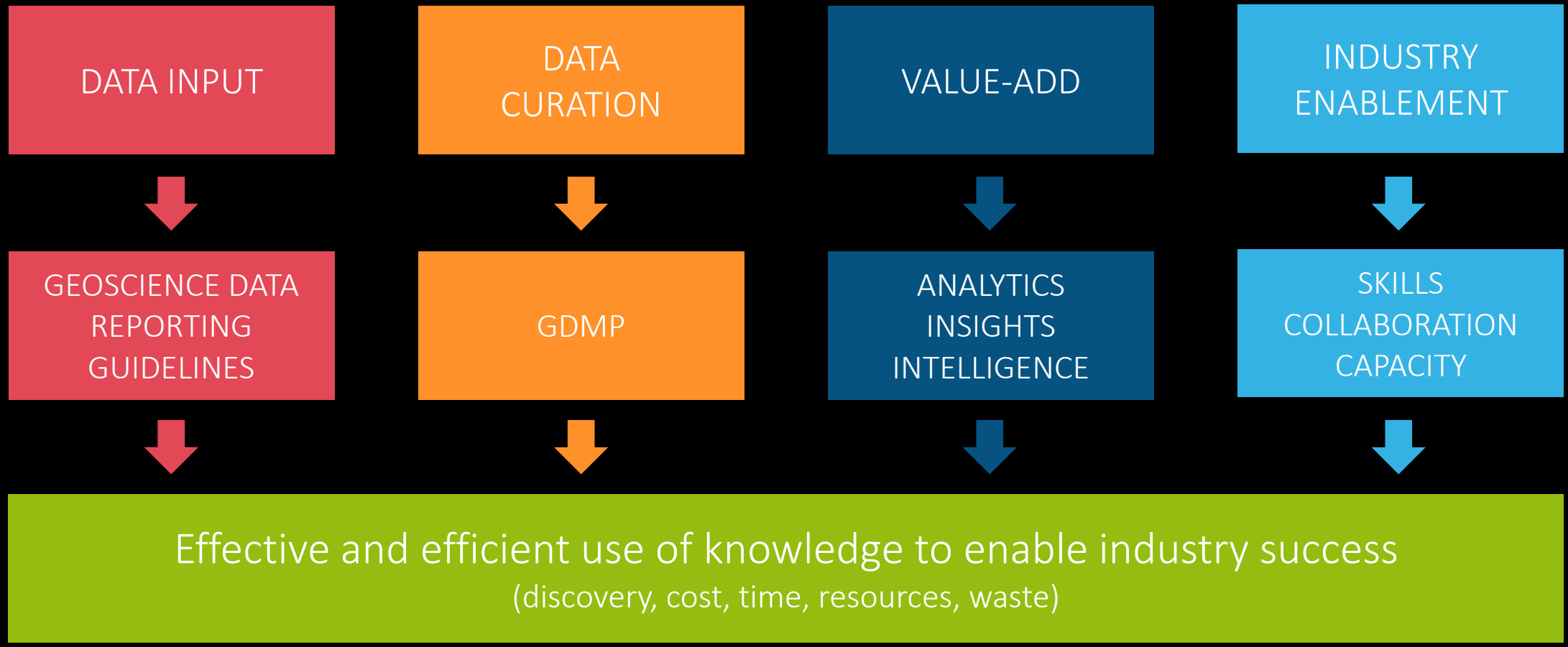
GSQ MISSION :

Focus our skills, technology and data to enable industry success.

GDMP – Data for Discovery

GDMP supports the GSQ vision to ‘enable industry success’ by making every possible piece of GSQ custodial data available.

Data-Driven Exploration



What are the current challenges?

DISCOVERABILITY	It is difficult to find all the geoscience information that exists in an area
ACCESSIBILITY	Industry can’t access all of GSQ’s geoscience data
QUALITY	Data quality is highly variable
USABILITY	We can’t unlock the value of our data
COST	Cost of maintaining multiple systems is unsustainable

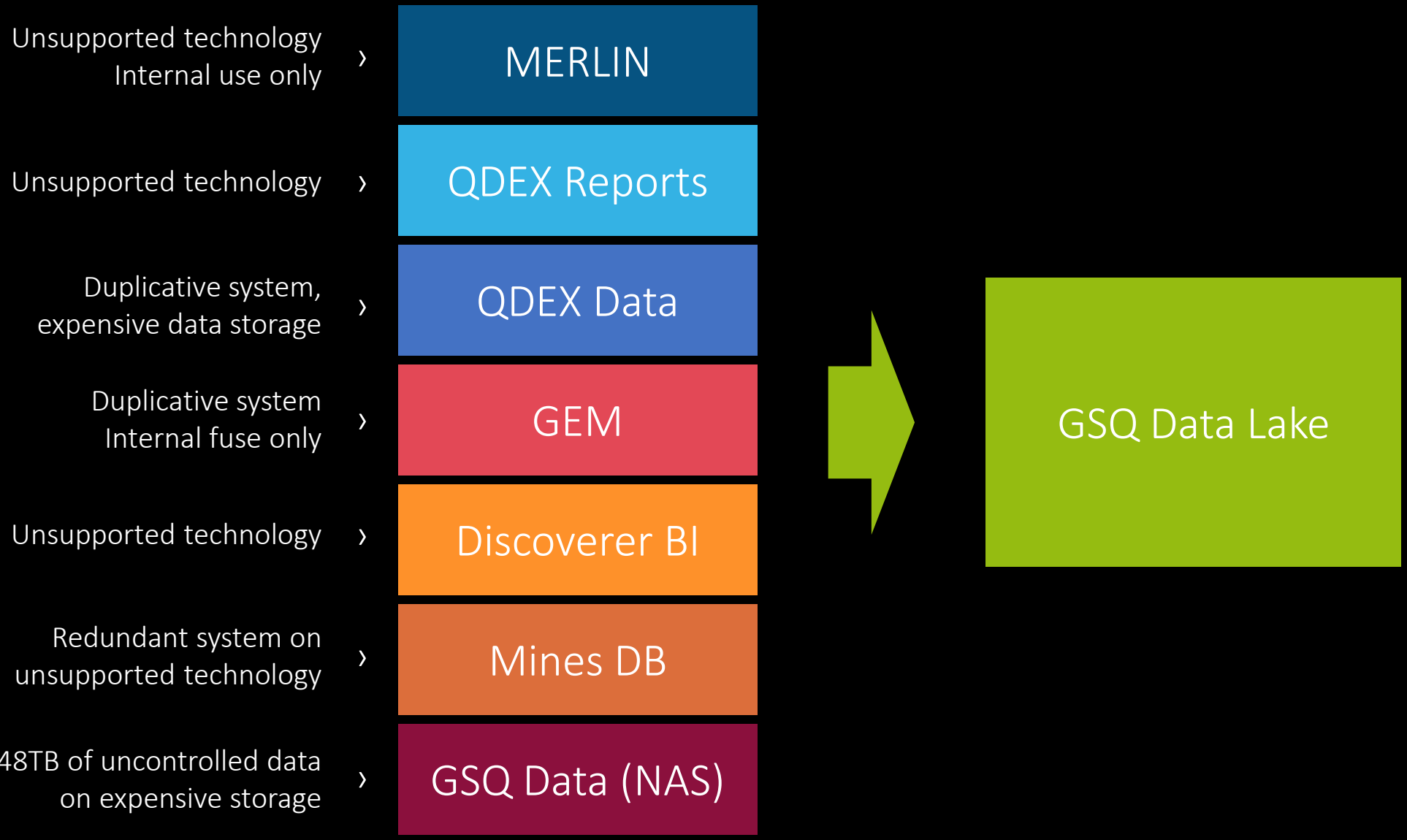
Data Management Lifecycle

CURRENT STATE

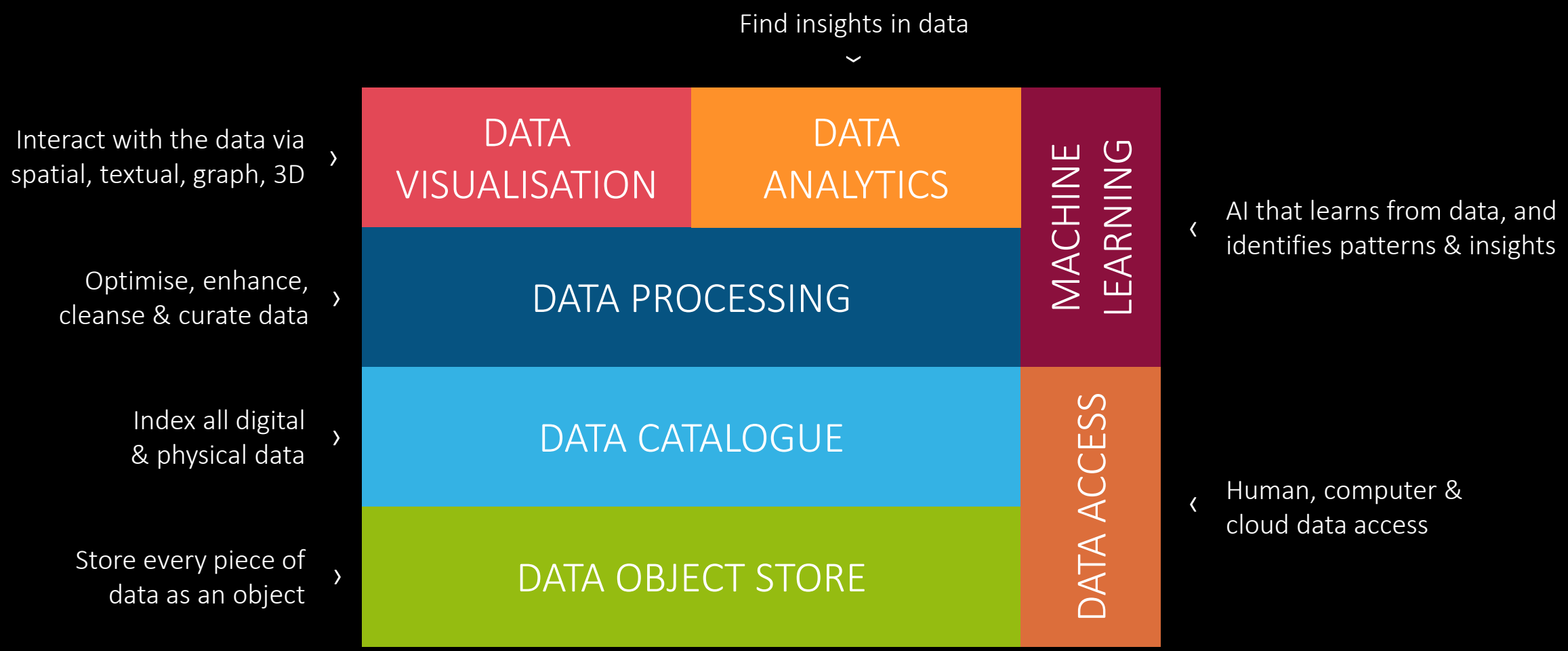
RECEIVE/ GENERATE	VALIDATE	STORE	CATALOGUE	DISCOVER	EXTRACT/ ANALYSE
Data unstructured & locked in PDF	Poor data quality with minimal QA	Multiple, fragmented expensive data stores	Multiple data catalogues	Low volume of data available to industry	Manual extraction, variable analysis
✘	✘	✘	✘	✘	✘
✓ Data itemised & well structured	✓ Data validated at point of entry	✓ Single, low-cost, flexible data store	✓ Single catalogue of all data	✓ >90% of open data available to industry	✓ APIs, open integration, advanced analytics

FUTURE STATE

System risk, cost, duplication



Data Lake Concepts



Data Catalogue Concepts

DATA CATALOGUE

A single catalogue of all data – digital, physical, federated

STANDARDISED DATA SCHEMAS

- A DCAT2 master schema to describe all data objects
- Extensible geoscience data schemas
- Schema validations

CONTROLLED VOCABULARIES

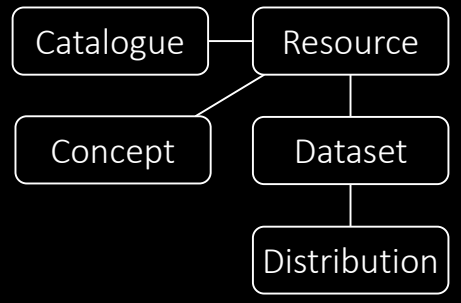
- A standard way to describe objects & their relationships
- Covers reference data, lookup lists, master data
- Preferred and variant terms

PERSISTENT IDENTIFIERS

- Globally unique identifiers
- Auto-provisioned e.g. IGSN for samples
- Alternate identifiers for historical data identities

LINKED DATA

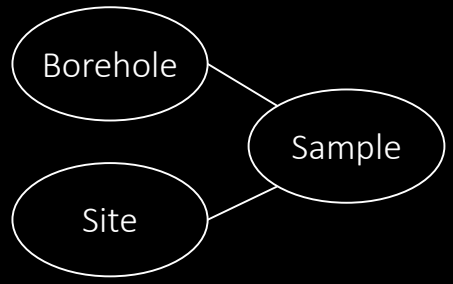
- Connects related data that wasn't previously linked
- Subject-predicate-object
- Human and machine-readable



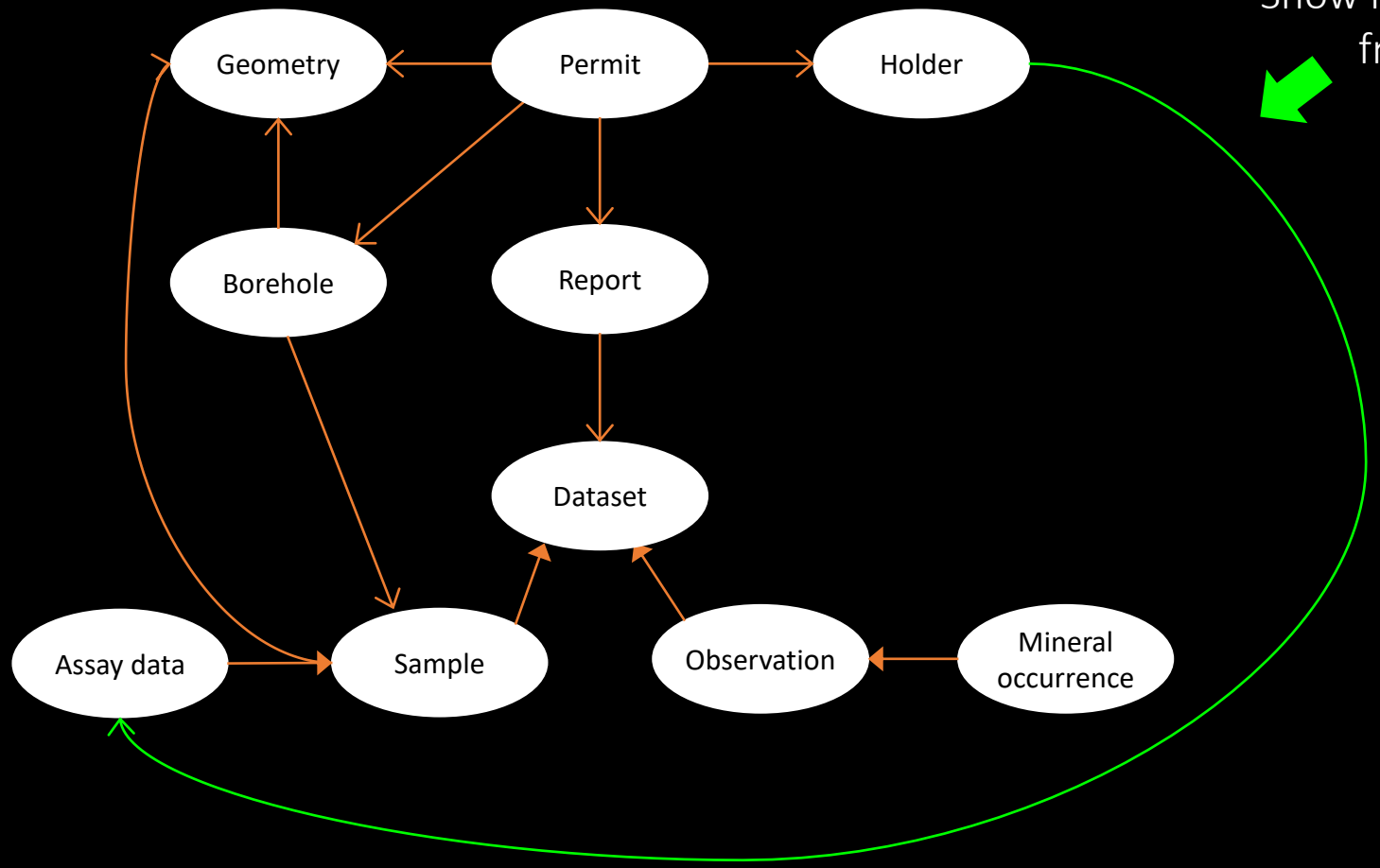
Vocabulary: Rock Types

- igneous
  - Adakite
  - Alkali feldspar granite
  - Basalt
    - 'A'a
    - Pahoehoe

//pid.geoscience.gov.au/sample/QG123



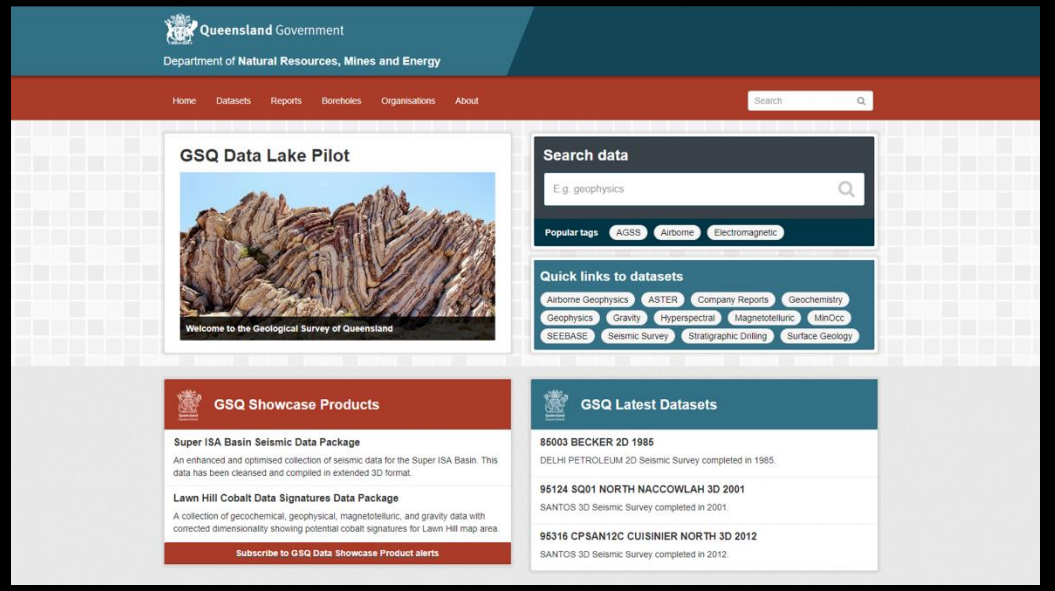
Linked Data



We can infer relationships:
Show me all the assay data from company x

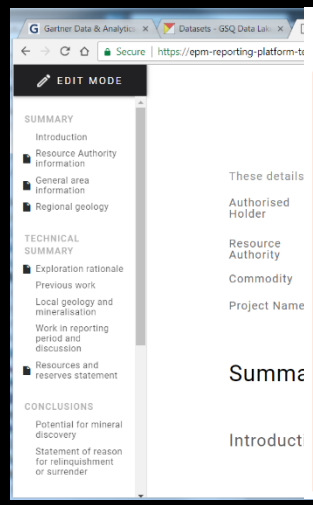
GDMP Pilot

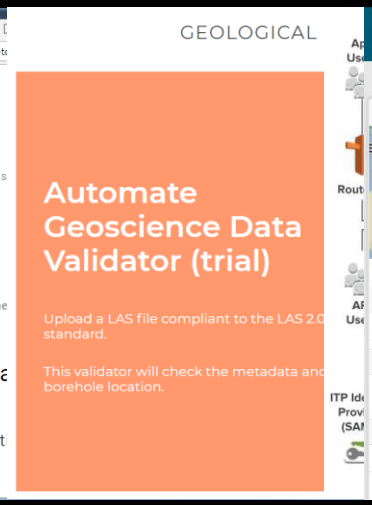
- GDMP industry pilot will be released June 30th.
- Focussed on seismic and geochemistry data
  - 800+ seismic surveys
  - Complete geochemistry database
- Users will be able to search for and download data
- Great feedback opportunity!

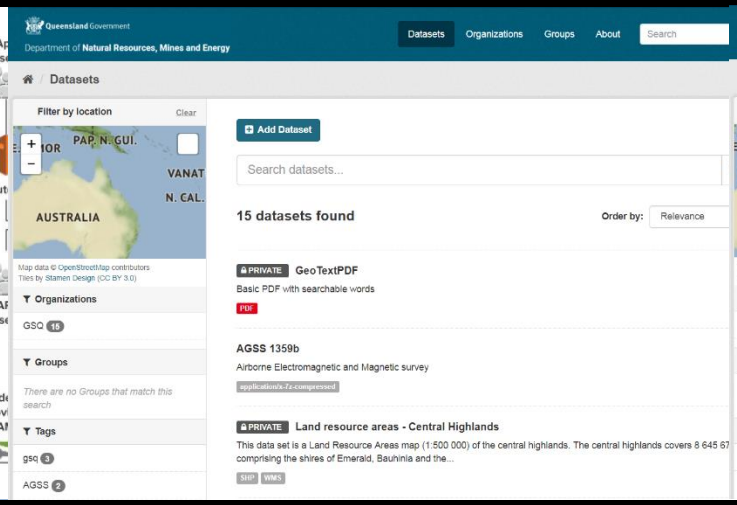


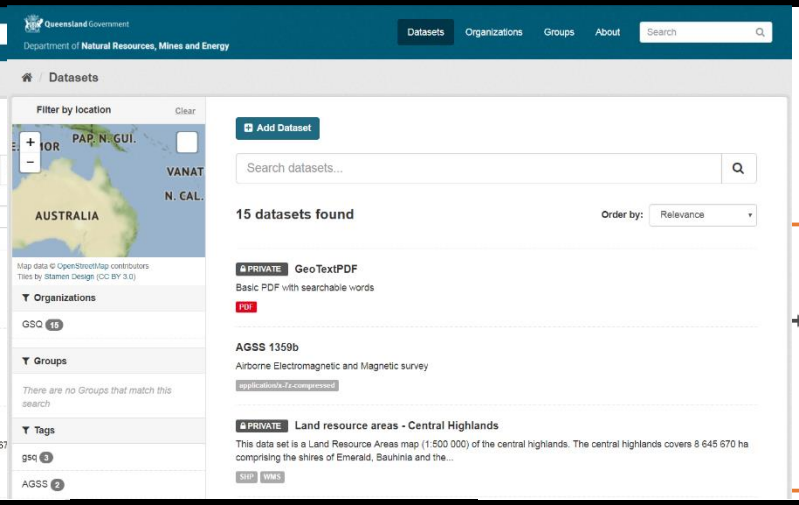
Pilot progress to date

RECEIVE/ GENERATE	VALIDATE	STORE	CATALOGUE	DISCOVER	EXTRACT/ ANALYSE
EPM report form prototype	LAS file validator	Pilot data lake	CKAN data catalogue Vocabulary manager	Data catalogue	FTP transfer File gateway









Why is GDMP different?

As-Is		To-Be
Variety of geoscience subjects	→	Standard representation of concepts
Many different data structures	→	Linked data structures – standards based
Diverse terminology	→	Controlled vocabularies
Data quality issues	→	Data quality metrics, quality assurance
Data trust issues	→	Data provenance recorded
Multiple metadata repositories	→	One metadata repository
Multiple data stores	→	One data store
Database primary keys, composite keys	→	Universally unique persistent identifiers
Complex database for metadata & data	→	Metadata in database, data in object store
Closed system inhibits data sharing	→	Open system, federated search
Designed for humans	→	Designed for humans and computers

Why is GDMP technically different?

Database/Data Warehouse

- Expensive data storage
- Technical skill to create data relationships
- Schema on write
- Vendor lock-in system
- Can store structured data

Data Lake

- Low-cost, high volume data storage
- Can infer data relationships
- Schema on read
- Mostly vendor-neutral system
- Can store structured, semi-structured & unstructured data

It’s all about data discovery!

Every data object

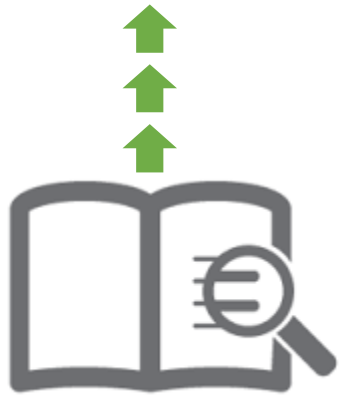
Is described by metadata

Metadata is validated

Against a controlled vocabulary

And a data store with integrity

For easy data discovery



Resulting in a high quality data catalogue

Opportunity statement

Optimise the geoscience data ecosystem
(*people, process and technology*)
to unlock the full value of data
to enable exploration success