Data from Drawers: Securing, mobilising and interrogating National Research Collections data

Andrew Young – Director, National Research Collections Australia (NRCA)

Australia: a mega-diverse continent

Australia has:

- A lot of biodiversity
  - 8% of the Earth's species
- Unique biodiversity
  - 70%+ endemic
- Valuable biodiversity
  - soybean, cotton, sorghum, macadamia, acacias, eucalypts

The challenge and opportunity is to:

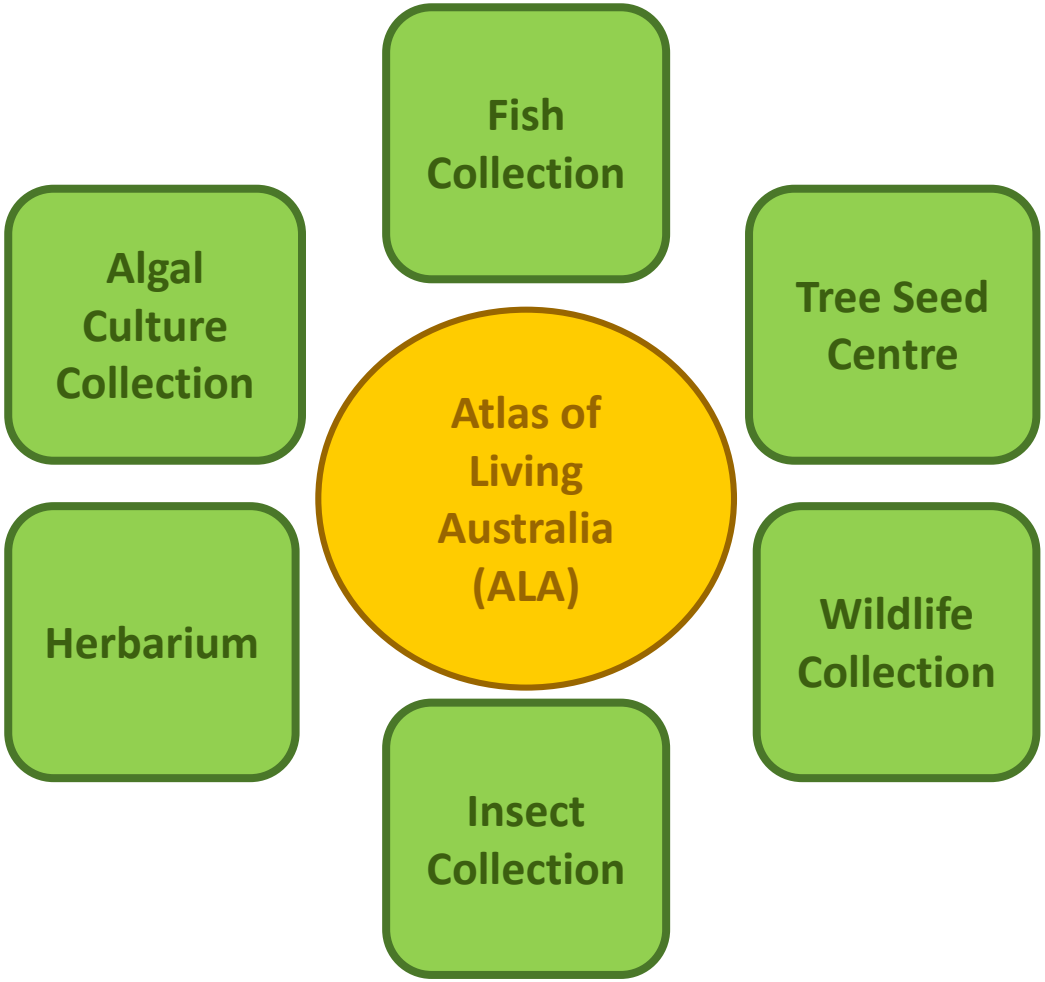
- Manage biodiversity for conservation and ecosystems services
  - species decline, Convention on Biological Diversity
- Exploit biological assets for industry
  - food, fibre, medicines, novel compounds

NRCA Mission

- National Research Collections Australia (NRCA) is a world-class "*science-ready*" collections research facility
- It discovers, documents, describes and explores Australia's biodiversity
- NRCA delivers digital data and science to inform the conservation and use of Australia's unique biological assets

What is NRCA?

- Six national biological collections
- 15+ million specimens
- 200 year time-series (1780)
- Atlas of Living Australia (ALA) web-based digital delivery and analysis capability



Soil Collection

Crop germplasm Collection

What is in NRCA?

- Physical specimens
  - whole organisms, skins, tissues samples, DNA samples
- Living collections
  - cultures, seed banks, seed orchards
- Digital specimens
  - sounds, photographs, X ray images, DNA sequences
- Contextual data
  - Location, site descriptions, species associations
- Unique $1+ billion research asset

Data challenges

1. **SECURE:** Integrated management of the data associated with the 15 million+ specimens e.g. single data system

2. **IMPROVE:** Increase the research value of collections through addition of new data layers e.g. metagenomics

3. **MOBILISE:** Digitize the collections for online data delivery e.g. specimen data, images, sounds

4. **ENABLE:** Online data delivery, visualization and analysis tools e.g. Atlas of Living Australia

1. SECURE: Integrated data management

Currently each collection has its own database:

- 5/6 are bespoke
- Only one is run by IMT
- Inefficient, ineffective and vulnerable…

Data management challenge – a single system:

- 15+ million specimens x 30-40 fields = 500 000 000 pieces of data
- Links to field books, living collections, nomenclature, associated samples (e.g seeds, tissues, DNA samples, sounds)
- Loans (30 000 - 40 000pa) and curation
- Room for future expansion (30 000+ pa)
- New data layers e.g genomes, images
- Biologically intuitive interface
- Seamless data delivery to the ALA

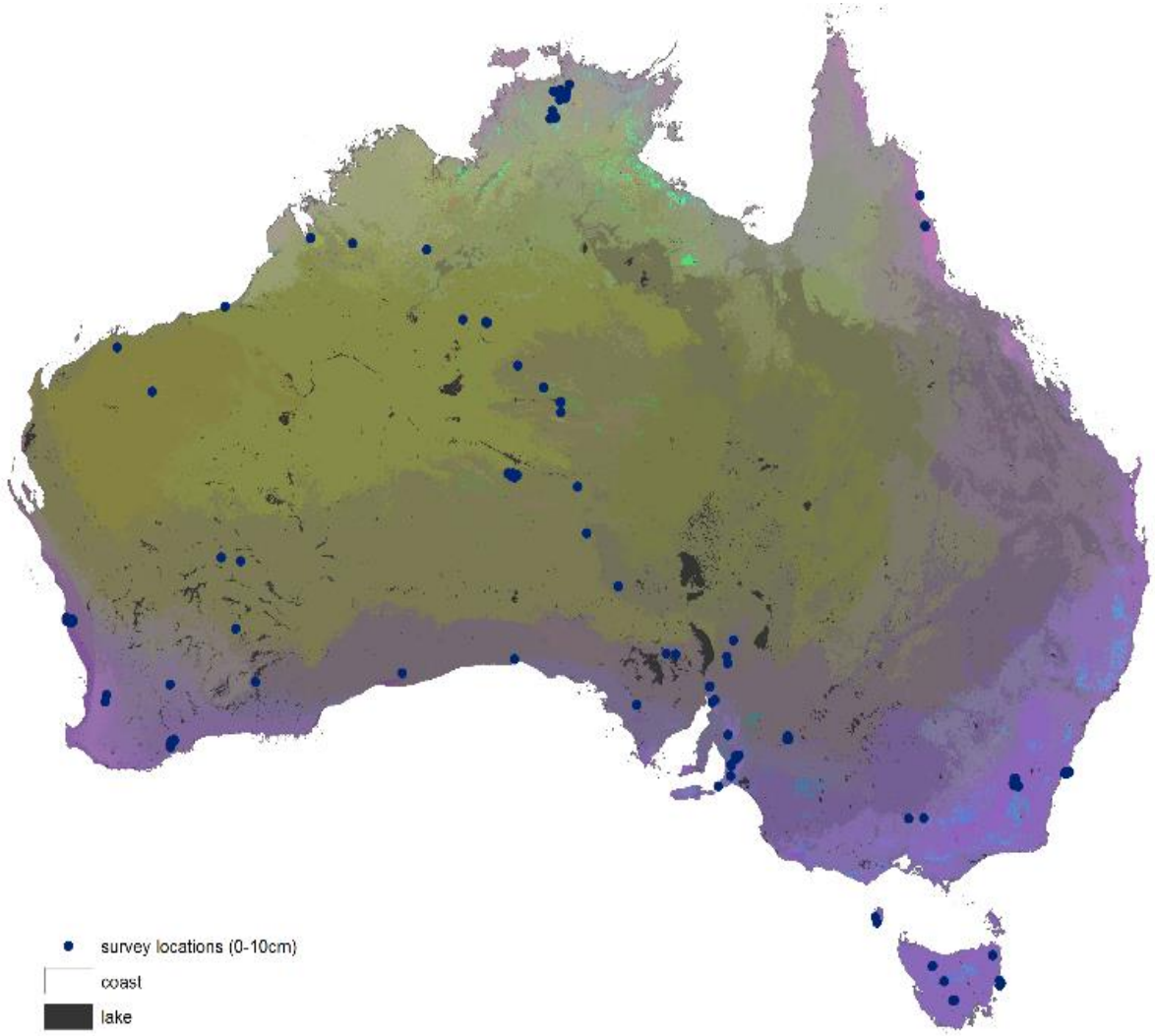
Collective Access

-Open source
-Thin client
-Fits IMT architecture
-Good functionality

2. IMPROVE: New data layers - metagenomics

- Soil microbes are ecosystem engineers
- BASE: A Genomic “National Framework Dataset”
- Bacteria, fungi and archea
- 1200 samples sequenced to 400-500,000 depth = 0.5 billion+ 400-800bp sequences
- 50+ physical variables
- Input, store, vizualise and analyze against other data layers e.g. vegetation, climate, new soil physical map
- Continental-scale predictive models of soil community structure and function

BASE Fungal diversity map 0-10cm horizon



3. MOBILIZE: Digitization

WHY DIGITIZE?
Secure – digital copy
Mobilise – other science users e.g. biosecurity
Expose – crowd source databasing


Picturae herbarium digi-street

- **Phase I:** Rapid digitization of 2-3 million specimens
- **Phase II:** Introduce digitization to current workflows - “born digital”

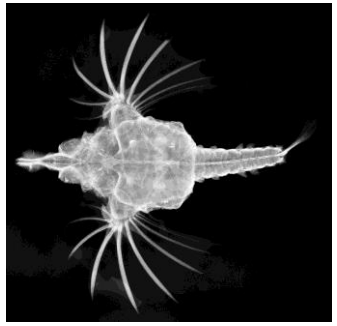
KEY ISSUES: Data volume (storage), prioritization & technologies

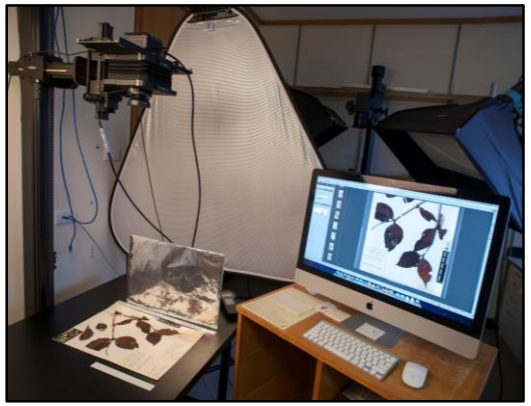
Last 5 years
Physical:
- 40 000+ loans a year

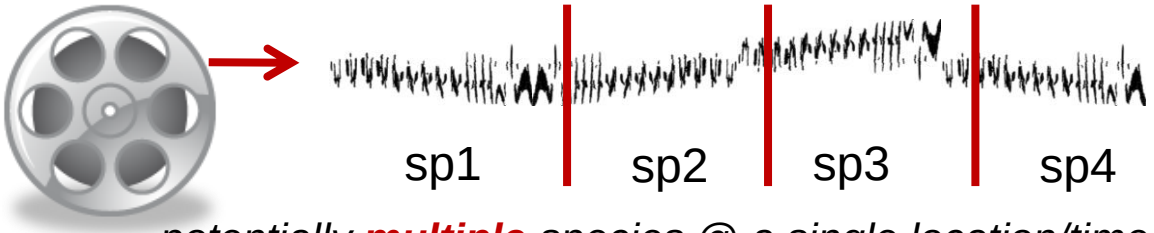
Digital ALA:
- 17 000 users a week
- 1.2 billion downloads a year

Current projects

- **ANIC** whole-drawer insect images
- **ANFC** digital radiography
- **ANH** Global Plants Initiative type specimen project
- **ANWC** bird sounds

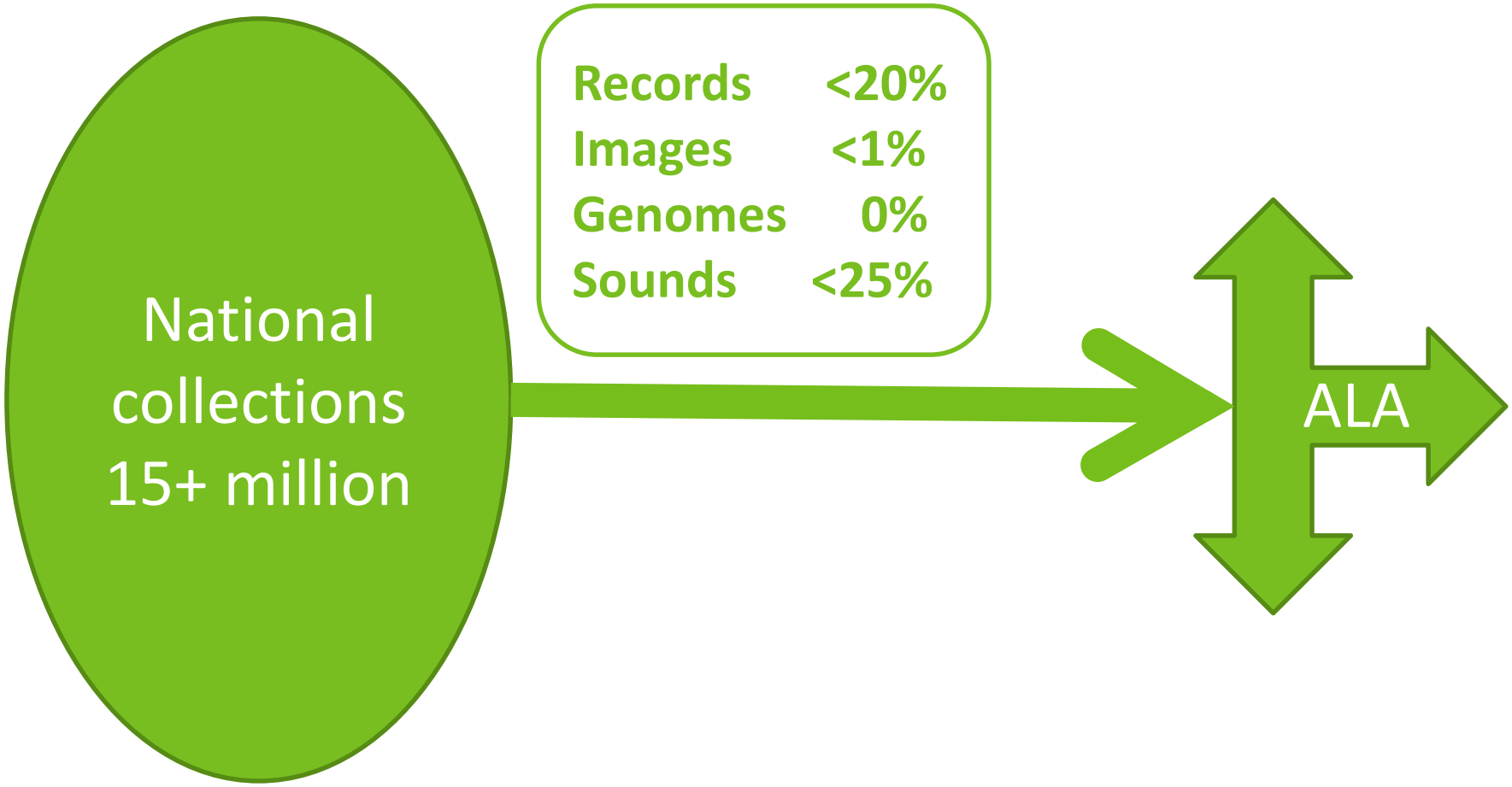






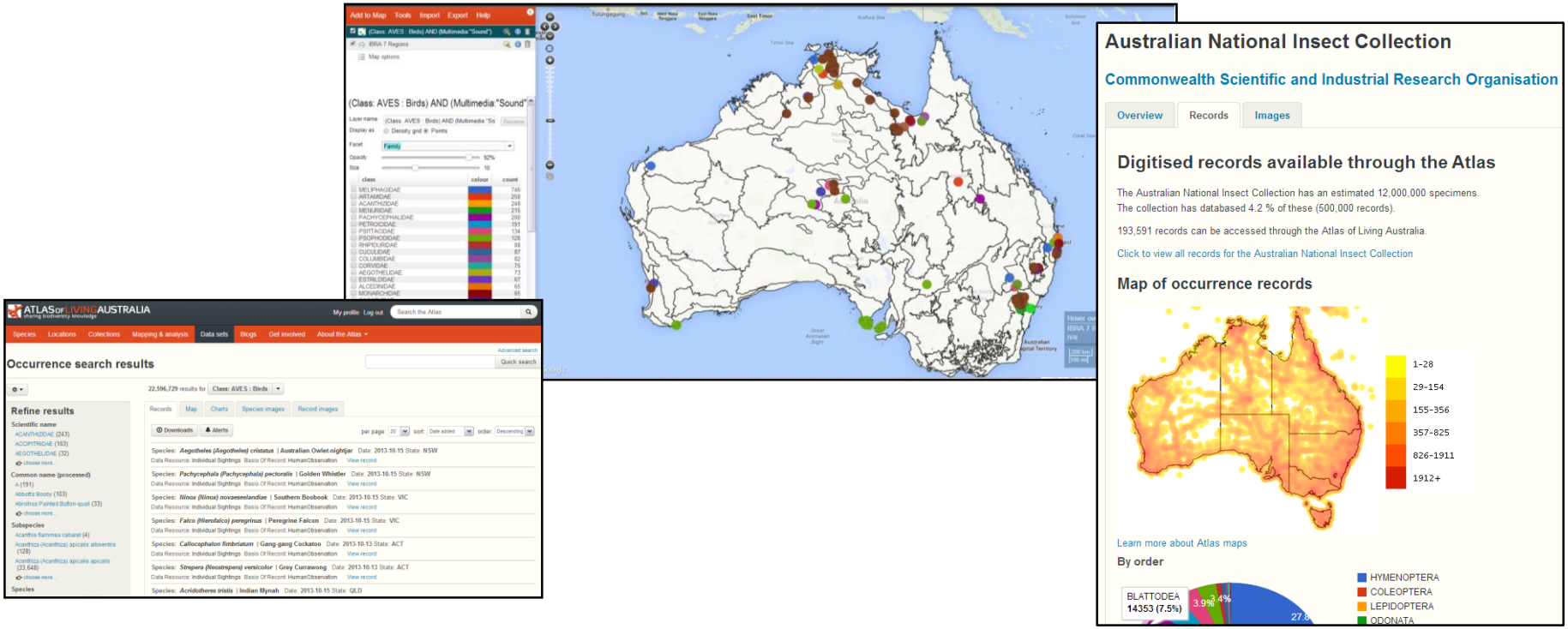
*potentially **multiple** species @ a single location/time*

Digitization challenge



4. ENABLE - ALA

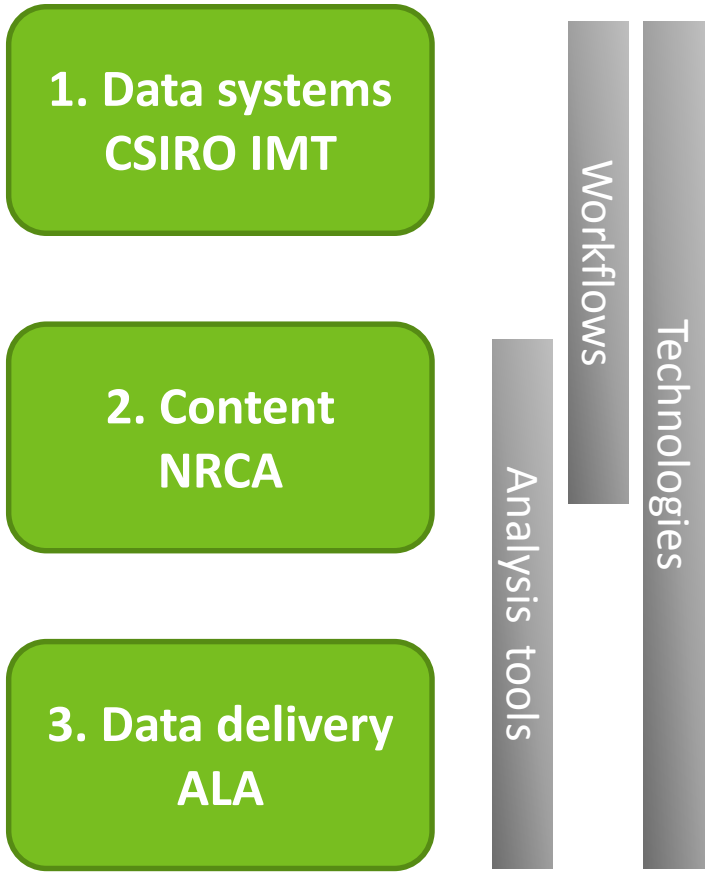
- Access to data in context of other data
- Visualization tools
- Analysis tools



NRCA digitization strategy

A work in progress….

- Data systems and storage
- Digitization technologies
  - sounds, images, sequences etc…
- Workflows
- Prioritization of specimens
  - types, rare and threatened, biosecurity, degrading data (e.g. sound tapes), user demand
- Analysis and data manipulation tools



Emerging challenges

Technical

- Extracting specimen data from entomological collections
- Genomic data
  - format and volume

Strategic

- Prioritization
- Selling the digitization value proposition
- Tracking impact not the downloads

Occurrence downloads by reason

Scientific research	**47,701 events**	1.22B records
Ecological research	**21,442 events**	528.46M records
Conservation management/planning	**7,234 events**	656.26M records
Education	**7,209 events**	114.22M records
Environmental impact, site assessment	**4,008 events**	124.98M records
Systematic research	**1,630 events**	83.14M records
Other scientific research	**620 events**	5.38M records
Collection management	**490 events**	45.04M records
Biosecurity management, planning	**428 events**	284.06M records
Other	**14,965 events**	198.15M records
TOTAL	**105,727 events**	3.26B records

more/less...

NRCA impact

Basic → **Applied**

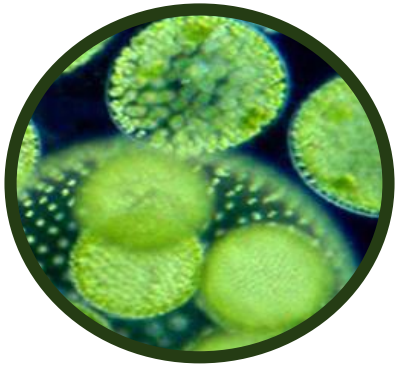
Environment ↓ **Industry**

Insect evolution

Re-vegetation

Marine reserves

Weed control



Bio-prospecting

Biosecurity

Thank you

Andrew Young – Director, National Research Collections Australia