The taxonomy and biosystematics decadal plan 2018–2027

Strategic Action 2.1: Building a comprehensive, integrated, accessible identification service for Australian organisms

Taxonomy and biosystematics decadal plan 2018–2027

Strategic action 2.1: Building a comprehensive, integrated, accessible identification service for Australian organisms

An implementation plan

Taxonomy Australia

November 2018

TAXONOMY AUSTRALIA
Discovering our biodiversity

Australian Academy of Science

Contents

Preface for DAWR report

This implementation plan was developed with funding from the Commonwealth Department of Agriculture and Water Resources (DAWR), under a contract awarded in June 2018 ("*Supporting a Coordinated Strategy for Taxonomy and Biosystematics to Improve National Diagnostics Capability in Biosecurity-related Taxonomy*") to the Australian Academy of Science.

The decadal plan for taxonomy and biosystematics, which this implementation plan underpins, deals with the taxonomy and biosystematics of all Australian biodiversity, with outcomes relevant to a wide range of activities including science, conservation, biosecurity, bioprospecting, resource management, citizen science etc.

The implementation plan has been developed to cover this broad scope and does not focus specifically on biosecurity. However, the diagnostics capability envisioned here has very direct relevance to biosecurity, in three ways.

Firstly, the integrated capability mapped out here will directly improve biosecurity diagnostics and provide a step change in the capacity of biosecurity diagnosticians to provide timely (including time-critical) and accurate identifications in the field and the laboratory. This in turn will lead to more effective management strategies both at points of entry and for the monitoring, control and potential eradication of pests within Australia.

Secondly, many aspects of biosecurity require a sound diagnostics capability for taxa that do not in themselves constitute a direct or current biosecurity threat. Examples include diagnosing native species that are closely related to taxa of biosecurity concern, quickly determining whether a specimen is native and low risk or needs to be diagnosed further, and building diagnostics capabilities for native taxonomic groups that may emerge as biosecurity threats in the future.

Thirdly, the level of integration that is key to the system proposed here will enable both effective identification and careful control on notifications of trade-sensitive taxa, neither of which are easy to achieve with the current *ad hoc* system of identification tools.

Building a system capable of providing a comprehensive, integrated, accessible identification service is a first step in its implementation. Populating the system with diagnostic data for taxa is a second step. Prioritising biosecurity-related taxa (including priority non-Australian taxa) in this second step will ensure that the service has maximum utility for biosecurity diagnostics early in its development, with extension to the remainder of Australia's biodiversity occurring in parallel and subsequently.

DRAFT

Executive Summary

Identifying (diagnosing) organisms is a key task for many people and organisations, with important use cases in biosecurity, conservation assessment and planning, taxonomy and biosystematics, ecology and other biological sciences, and in enabling members of the general community to identify and understand organisms they see and that interest them.

However, tools that enable identifications of organisms are very varied, are widely scattered, have been developed largely *ad hoc*, are poorly integrated and are often poorly maintained.

The decadal plan for taxonomy and biosystematics, developed by the Australian and New Zealand taxonomy and biosystematics sectors under the auspices of the Australian Academy of Science and Royal Society Te Apārangi, recognised this weakness and listed amongst its strategic actions the development in the next decade of a comprehensive, flexible, integrated, accessible service for identification of Australian organisms.

This implementation plan provides a roadmap for realising this strategic action.

It establishes a general framework for understanding identification and diagnosis, describes the main modes of identification, describes a modular architecture of independent but interoperable online tools that can connect and integrate these modes, and establishes tasks, timelines, responsibilities and costs for building the system described, a system that would deliver a significantly enhanced capability for biodiversity diagnostics in Australia.

Key to the vision is the integration of identification tools that use different and complementary modes for identification, including trait-matching (using tree-based and matrix-based identification keys), image-matching, DNA sequence identification, machine learning and geographic scoping.

This will allow, for example, an identification to commence with determination that a specimen belongs to a given genus using a DNA sequence, with the identification then passed to a trait-based key that reduces the set of potential species to a handful, followed by the assembly of a set of diagnostic images to provide an identification to a single species.

A shared, foundational taxonomic framework, attention to standards, and the development of web services that will enable interoperability will allow this.

At the same time, the integrated system described here will allow identification tools to be kept up-to-date in the face of changing taxonomies, the curation and maintenance of existing identification tools, and the efficient development of new identification capability.

With sufficient support and resourcing, the system described here could be built over a 2-year development path with an estimated 9FTE for system design, development and initial management (Table 1).

As well as initial development, the success of the enhanced diagnostics capability described here will depend on ongoing strategic development and maintenance. This could be best achieved by the establishment of a dedicated, multi-agency program, provisionally called the Australian National Biodiversity Diagnostics Program.

DRAFT

Table 1. Resourcing estimates for development and initial management of an enhanced biodiversity diagnostics capability in Australia (see Appendix 1 for details)

Task	Potential Partners	Estimated resources
1: ***Establish an Australian National Biodiversity Diagnostics (ANBD) program to coordinate implementation of this plan.***	Department of Agriculture and Water Resources, Atlas of Living Australia, CSIRO, Subcommittee for Plant Health Diagnostics, Plant Health Australia, State and Territory herbaria and museums, Taxonomy Australia	$160,000 per annum for 3 years to appoint and maintain a National Coordinator and support activities of the Steering Group
Task 2: ***Develop an integrated trait-based matching module***	Taxonomy Australia, Department of Agriculture and Water Resources, ABRS, State and Territory herbaria and museums	$110,000 p.a. for 1 year to develop the module and $110.000 p.a. for two years to upload legacy keys
Task 3: ***Initiate a campaign for strategic capture of diagnostic images of Australian taxa***	Department of Agriculture and Water Resources, Atlas of Living Australia, Australian National Botanic Gardens, State and Territory herbaria and museums, Taxonomy Australia	$110,000 p.a. for 1 year to develop the module and 110,000 p.a. for 2 years to aggregate content and engage citizen scientists
Task 4: ***Develop an annotations layer and identification module that will enable efficient and effective DNA-based identification of Australian organisms***	BioPlatforms Australia, Atlas of Living Australia, State and Territory museums and herbaria, Taxonomy Australia	$110,000 per annum for 2 years
Task 5: ***Assess and develop machine learning approaches to biodiversity diagnostics***	Universities in Australia and overseas	No estimate
Task 6: ***Develop an overarching integration module to allow interoperability between diagnostics modules***	Department of Agriculture and Water Resources, Taxonomy Australia	$110,000 per annum for 1 year
Total		**$1.47M over 3 years**

DRAFT

Context

The decadal plan for taxonomy and biosystematics in Australia and New Zealand 2018–2027[1] outlines the foundational importance, impact and relevance of species discovery and other aspects of taxonomy and biosystematics for science, society, industry and government, and presents an agreed vision for the sector for the next decade.

The vision focuses on creating a step change in species discovery and biodiversity documentation, and through this a step change in understanding, protecting, managing and sustainably using Australia's and New Zealand's biodiversity.

The vision is supported by 22 strategic actions based on six key initiatives: *accelerating discovery*; *enhancing services*; *engaging with Indigenous knowledge*; *improving infrastructure*; *educating for the future*; and *building strategic capabilities*. These strategic actions were developed through extensive community consultations, both within the taxonomy and biosystematics sector and with key stakeholders.

The strategic actions in the decadal plan are outlined briefly but are not developed or discussed in detail: the plan provides a vision and destination, but not a roadmap to build the vision and get to the destination.

For that reason, a series of implementation plans is being developed to underpin the decadal plan. Each implementation plan will be relevant to one or more of the decadal plan's strategic actions and will set out in detail the strategic action's outcomes and objectives, analyse its dependencies and risks, and establish a timeline and budget for its implementation.

Because of substantial governmental, structural, biological and social differences between Australia and New Zealand, parallel and complementary implementation plans will be developed for each country.

This document provides an Australian implementation plan for the decadal plan's Strategic Action 2.1:

> *We will create a comprehensive, flexible, integrated, accessible service for identification of Australian and New Zealand organisms, based on DNA sequences, morphology, and images.*

This strategic action is part of the key initiative to enhance services for end-users of the knowledge provided by the taxonomy and biosystematics community. It will be directly relevant to any individual or agency with a need to identify an organism in Australia. This includes users in biosecurity, agriculture, fisheries, aquaculture, conservation, resource industries, ecology and other biological sciences, and members of the public who identify organisms for interest or enjoyment.

[1] https://www.science.org.au/support/analysis/decadal-plans-science/diversity-decadal-plan-taxonomy

Background

The timely (and sometimes time-critical) identification of organisms is a key task for many people. It is also a key concern for taxonomy and biosystematics, the disciplines that resolve, delimit, name and classify species and other taxa[2]; such resolution and delimitation is of little value if others are unable to identify the taxa so resolved.

Identification is any process that enables someone to determine the correct name of a specimen or individual of an organism. Simple cases of identification include birdwatchers recognising a bird species by plumage or call, or a wildflower enthusiast identifying a plant by comparing it with a photograph in a field guide. More complex forms of identification use increasingly complex tools, from printed and electronic identification keys to gene sequencing and machine learning.

In general terms, identification is a reduction in potential taxonomic scope for the specimen or individual organism being identified. Before identification, the taxon that a specimen or organism belongs to is indeterminate; after identification the specimen or organism is determined to belong to one taxon, at an appropriate rank.

Identification is usually a nested, step-wise process, which uses the taxonomic hierarchy developed by taxonomy and biosystematics. Identifying an insect found in a consignment of goods at a port in Australia to the level of family or genus may be sufficient to determine its biosecurity threat. Equally, an identification that determines the family the insect belongs to may necessitate a subsequent identification step to determine its genus or species if biosecurity risk varies within the family.

There are many available methods or modes of identification, ranging from gestalt (as when an experienced person can immediately recognise an organism and correctly place it in a taxon), to complex analytical procedures. The most common modes are as follows:

Mode	Process
Gestalt	A person with experience of a taxon instantly recognises it by comparison with prior experience or knowledge (as when a birdwatcher instantly recognises a bird, or a taxonomic expert instantly assigns a specimen to a known taxon).
Visual pattern-matching	A person matches the overall features of a specimen or organism to a named image or photograph in e.g. a field guide.
Tree-based trait-matching	A person follows a guide structured as a tree of traits (such as a printed dichotomous identification key), progressively narrowing the taxonomic scope while proceeding through the tree.

[2] A *taxon* (pl. *taxa*) is any formally classified and named unit of biodiversity. Species, subspecies, genera, families, orders, phyla etc. are all taxa.

DRAFT

Matrix based trait-matching	A person describes the specimen by choosing matching traits in a system (usually computer-based) that progressively reduces the taxonomic scope by filtering out taxa that do not match the set of chosen traits.
Genome matching	A computer algorithm compares a gene sequence for a specimen to a library of named gene sequences, and determines the most similar sequence(s).
Machine learning	A trained neural network compares an image of a specimen with parameters derived from images in its training set, and determines the closest-matched taxon or taxa.
Geographic scoping	While not an identification mode in its own right, geographic scoping cuts across all other modes; if a specimen comes from a known location, and if the set of taxa known or predicted to occur at that location can be reasonably predicted, this can be used to reduce the initial taxon set for the other modes.

All these modes of identification are used in Australia.

Gestalt identification

Gestalt identification involves the recognition of a specimen as a taxon without the necessity for analysis or external aides. Examples of gestalt identification include someone identifying a beetle as a beetle, without necessarily understanding the critical characteristics of Coleoptera, or an experienced taxonomist accurately placing a specimen in its correct species based on a deep knowledge of the taxonomic group to which it belongs.

For people who need to identify organisms, gestalt identification is a goal, and a skill that builds over time. It is often highly efficient – a good identifier can accurately place a specimen into a taxon with which they are familiar quickly and with minimal apparent effort.

Importantly, accuracy and breadth of ability for gestalt identifications is built by accruing diagnostic experience. This needs solid training, which may be obtained by close taxonomic study of the group in question, or by repeated identifications using other, more analytical, diagnostic modes. For example, an identifier who uses traditional dichotomous keys repeatedly is more likely to build a strong foundation for accurate gestalt identifications than one who performs fewer identifications or uses modes of identification that do not require the close attention to detail that builds the knowledge base required for accurate gestalt identification.

Given the efficiency and efficacy of gestalt identification, an important long-term question is the effect of a current move away from traditional, analytical identification and taxonomic

DRAFT

techniques, such as dichotomous keys, “apprenticeships” with other skilled identifiers, and taxonomic revisionary work, to other methods such as genome sequencing and machine learning (see below). Over-dependence on these may improve identification capabilities in the short term but result in a long-term loss of expertise in gestalt identification.

Gestalt identification to a broad taxonomic level is a default mode for most identifications (that is, most people can correctly identify a bird, insect etc). For more rigorous identifications, it is usually the starting point, allowing an identifier to commence a more analytical identification at a relatively low taxonomic level rather than starting with “all life”.

Visual pattern-matching

Identification by visual pattern matching involves the comparison of a specimen with named images or specimens of taxa. Examples include the use of field guides, online image-banks, or direct comparison with reference collections.

Pattern-matching may be effective when distinctions between taxa are relatively clear-cut. It is likely to have a high misidentification rate when there are closely related taxa that differ by traits that are not obvious, particularly to an untrained eye.

In Australia, pattern-matching identification guides (field guides etc) are available for a limited number of taxonomic groups, mainly those of broad public interest such as birds and wildflowers. Such guides are usually taxon-complete only for relatively species-poor groups such as vertebrates.

Tree-based trait matching

Identification by tree-based trait matching involves the use of traditional, usually dichotomous keys. Pairs of descriptors (couplets of traits) in such keys are arranged in a hierarchical tree; a user navigates through the tree by choosing one lead (branch) at each couplet (node), thus navigating a path through the tree to a named taxon.

Tree-based trait matching is the most common analytical identification method and has a long history. Printed dichotomous keys are still used widely in taxonomic treatments and are a key resource for anyone who needs to go beyond simple pattern-matching for identifications.

Printed dichotomous keys, while effective and efficient, have two major drawbacks. Firstly, they are highly inflexible: a user is constrained to follow a path through the tree, and paths may be blocked by unanswerable couplets. Secondly, space constraints on printed keys mean that they often depend on precise and technical (“jargon”) language, and this limits their use outside domains of expertise.

These drawbacks are offset to some extent by a major advantage: repeated use of dichotomous keys for a given taxonomic group is an effective way to build the knowledge base to enable gestalt identifications. Dichotomous keys provide structured knowledge of key diagnostic traits within a group in a form that can be readily learnt and internalised.

In Australia, most dichotomous keys are printed and are widely dispersed. This provides a challenge for users: how to find the right key? An online repository and deployment

DRAFT

platform for dichotomous keys, KeyBase[3], has been developed by the Royal Botanic Gardens Melbourne, and contains over 8000 keys in a standard, searchable format. To date it has been used almost entirely for botanical keys, but is general in scope and could be used more widely.

Matrix-based trait-matching

Identification by matrix-based trait matching is a common computer-based identification mode, best exemplified by systems such as Lucid[4] and DELTA[5]. A traits × taxa matrix is constructed. Users of the system sequentially (but in any order) choose traits that match the specimen being identified. The system responds at each step with a list of taxa that match all (or a majority of) chosen traits. Choosing more traits progressively reduces the list of matching taxa until (potentially) only one remains and the identification is complete.

A significant advantage of matrix-based trait matching over tree-based trait-matching is that a user can choose any trait at any time, rather than being constrained by fixed pathways in a tree of traits. This often allow users to avoid the unanswerable-couplet problem.

In formal terms, this is achieved by increasing trait redundancy. In a tree-based trait key, each node of the tree minimally comprises a single trait (couplet); such nodes effectively have no redundancy (there are no alternative pathways). The matrix upon which matrix-based trait matching is performed, by contrast, has maximal redundancy; every trait is theoretically relevant to every taxon and a user who cannot address one trait can address any other instead.

While the high redundancy of matrix-based systems provides an effective solution to the unanswerable-couplet problem, it comes with two costs. Firstly, the extra redundancy requires more work to assemble the underlying matrix. Associated with this, there is a higher likelihood of errors (traits incorrectly assigned to taxa), particularly in traits that are not highly diagnostic and hence not well-known.

Both matrix-based and tree-based trait-matching modes of identification are widespread and popular with users.

Genome matching

The genome of a taxon is by definition highly diagnostic. With the ready availability of partial (and, increasingly, full) genome sequences, identification based on matching a sequence from an un-named organism against a library of sequences of named organisms is becoming increasingly mainstream.

The most common matching algorithm is BLAST[6], a freely available, fast, heuristic algorithm capable of efficiently finding close sequence matches in a library comprising millions or

[3] https://keybase.rbg.vic.gov.au/
[4] http://www.lucidcentral.com/
[5] https://www.delta-intkey.com/www/interactivekeys.htm
[6] Journal of Molecular Biology. 215 (3): 403–410. doi:10.1016/S0022-2836(05)80360-2. PMID 2231712.

hundreds of millions of named sequences. Being heuristic, BLAST is not guaranteed to find the closest match, but has been shown to perform well in most cases.

For identification, BLAST requires an input sequence (from the organism to be identified) and a library of named sequences. The two most commonly used libraries are GenBank[7], managed by the US National Center for Biotechnology Information (NCBI), a publicly accessible library of >200 million sequences, and the Barcode of Life Database (BOLD[8]), which includes >6 million barcode (short-marker) sequences. GenBank includes sequences from >450,000 species and infraspecies taxa, and BOLD includes barcodes from >280,000 species and infraspecies taxa, representing c. 10% and c. 5% of the estimated number of named species in the world respectively.

Two factors currently limit universal use of genome-matching. Firstly, c. 90% of named species are not yet represented by sequences in GenBank, BOLD, or other repositories (although coverage of taxa is highly uneven and for some taxonomic groups coverage is substantially better). Secondly, poor quality control and taxonomic curation in the early stages of development of GenBank means that many sequences are unvouchered and their names are incorrect or doubtful. For these two reasons, the names returned from a BLAST search of GenBank must be treated with caution. BOLD is more rigorously curated, including a requirement that contributed sequences must be vouchered. However, the short barcode sequences used in BOLD are useful for some but not all taxonomic groups, and barcoding is likely to be superseded by next-generation genome sequencing in the near future.

BLAST searches against GenBank and BOLD are likely to perform somewhat better at higher taxonomic ranks. Nearly 20% of genera are represented by one or more sequences in GenBank, and this percentage rises at higher taxonomic ranks. Thus, used with care, identification of an unknown specimen to family or genus level using a BLAST search of these databases could be expected to perform better than identification to species level. However, accumulation of sequences in GenBank is *ad hoc* (less so in BOLD), resulting in highly uneven coverage of taxa at every rank.

In addition to BLAST searches against the GenBank and BOLD libraries, a small number of more specialized services have operationalized genome-based identifications for small groups of organisms. Examples include a *Fusarium* identification tool developed by the Westerdijk Fungal Biodiversity Institute[9], a prokaryote identification tool developed by the commercial company Ribocon[10], and a partially completed tool for identifying plants of the Pilbara region using chloroplast sequences[11].

Machine learning

Machine learning is the fastest-growing field in computer science, driven by a combination of factors including the ready availability of fast, parallel-core computer hardware, the

[7] https://www.ncbi.nlm.nih.gov/genbank/
[8] http://www.barcodinglife.org/
[9] http://www.westerdijkinstitute.nl/Fusarium/
[10] http://jspecies.ribohost.com/jspeciesws/
[11] https://pilbseq.dbca.wa.gov.au/

development of deep convoluted neural networks, and substantial industry and government investment in machine learning technologies for commercial use cases such as facial recognition and driverless cars. The application of machine learning algorithms to species identification has also been accelerated by the ready availability of very large numbers of images of species, often provided by citizen scientists with digital cameras and smart phones.

Machine learning holds substantial promise for identifying images[12] of species in the field or specimens in the laboratory or in collections. However, machine learning is probably close to the peak of its "hype curve"[13], where expectations exceed likely real outcomes by a wide margin, particularly for these applications.

While many studies are testing the utility of, and refining algorithms for, biological identification by machine learning, few are probing its limits. A common assumption is either that there are no limits, or the rapid development of the technologies means that investigation of limits is meaningless or premature.

For example, the largest test yet of the capabilities of machine learning for identification of species[14] used as a test set photographs contributed by citizen scientists to the European citizen science project Pl@ntNet[15]. Such a test set is inherently weighted towards visually prominent and striking species that are likely to be noticed and photographed by Pl@ntNet contributors. While the test results from some of the competing systems are impressive (a reported >85% accuracy), this is likely to be inflated to some degree by the inherent bias in the test set. This is thus a real-world test for the identification of citizen science images, but not for the identification of species for other purposes.

Nevertheless, machine learning systems clearly have a demonstrated role in species identification in some situations, and are developing rapidly.

Geographic scoping

Geography is a powerful predictor of biodiversity: clearly, the set of taxa that occur in south-west Western Australia is very different from the set of taxa that occur in north-east Queensland. Thus, taking the broad definition of identification as any process that reduces the potential taxonomic scope for a specimen or individual organism being identified, using geography to limit the scope of an identification is a cross-cutting identification mode.

Geographic distributions are well-understood for some taxonomic groups, based on the aggregate of specimens in biodiversity collections and non-vouchered observation records where available. Programs such as the Atlas of Living Australia (ALA), Australasian Virtual Herbarium (AVH) and Online Zoological Collections of Australian Museums (OZCAM) provide relatively ready access to this information.

[12] While mostly used for identifying images, machine learning can also be used for identifying calls etc.
[13] https://www.ledgerinsights.com/gartner-blockchain-hype-cycle/
[14] http://publications.hevs.ch/index.php/attachments/single/1168
[15] https://identify.plantnet-project.org/

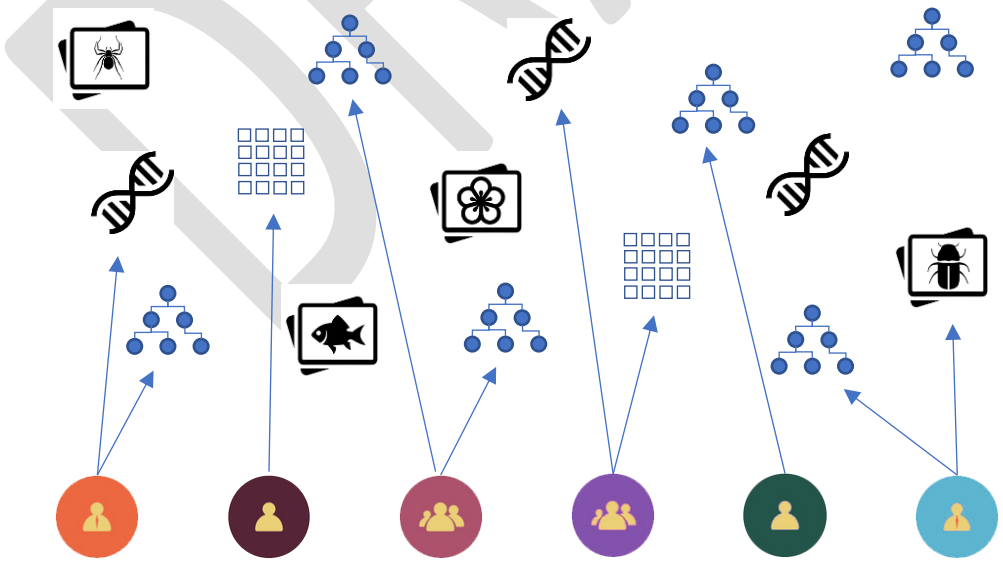
All can, at least theoretically, provide a checklist of species known or expected to occur at a given location, using either known records or distributions inferred from species distribution modelling. Such checklists can then be used as a starting point, or taxonomic scope, for other identification modes. However, limited databasing of many faunal collections in museums limits the effectiveness of geographic scoping for many taxonomic groups, particularly invertebrates.

The Issue

There is a very large number of identification resources (keys etc) in use in Australia for identifying species and other taxa, for a wide variety of purposes, spread over a range of identification modes. However, there is very little integration, either within or between these modes. This lack of integration has the following consequences:

1. Low discoverability of resources (a user may need an identification resource, but be unable to find an appropriate one)
2. Low accessibility of resources (a user may not be able to readily access a resource even if known)
3. Limited coordinated management of identification resources (resources are often developed *ad hoc*)
4. Limited ongoing curation and management of identification resources (including updating to new taxonomies, adding new species etc.)
5. Limited ability for users to use multiple resources in concert to make an identification

All these limit opportunities for timely (and sometimes time-critical) identifications, and increase the risk of inadequate, inaccurate, or insufficiently timely identifications.



The current situation. A wide range of disconnected, *ad hoc* identification resources, including trait-based keys, DNA diagnostics protocols and image reference libraries are

available, but are poorly discoverable for users, are not integrated, and may or may not be managed in the face of changing taxonomies.

Some modes (e.g. trait-based identification) are mature, highly operationalised and widely used across many taxonomic groups, while others (e.g. genome matching and machine learning) are undergoing rapid development and are currently used in only limited domains, often comprising experimental or test-bench systems with limited operational roll-out. The lack of a well-designed, strategic framework for developing these new systems is likely to limit their utility and relevance across all use cases.

Similarly, the lack of an integrated framework that accommodates all identification resources reduces efficiencies and limits options for long-term maintenance and management.

Some identification modes (e.g. image-based pattern-matching and machine learning) are limited by the lack of a broad-based, curated, accessible, authoritative image collection for Australian biodiversity. Very few taxonomic groups have well-curated image libraries. The Australian Plant Image Index (APII[16]), for example, is curated, accessible and authoritative, but focuses on only one taxonomic group (plants) and is not suitably broad-based – there are many diagnostic images of taxa available in other repositories that are not included in the APII. The wide dispersion of images, lack of clear standards for adequacy, and lack of ongoing curation of images in many repositories means that it is currently impossible to determine which Australian taxa (even amongst plants) have available adequate images and which do not.

Of course, taxonomic identification critically requires an effective underpinning taxonomic framework – the species (and other taxa) being identified and their names. Australia has well-developed mechanisms for maintaining nationally agreed taxonomies. The National Species Lists[17] and Australian Faunal Directory[18], managed by the Australian Biological Resources Study (ABRS[19]) provide nationally agreed checklists of Australian taxa in all groups of organisms, and to some extent a framework for managing synonymies and resolving older names to accepted names.

While identification resources remain widely scattered and highly decentralised, there are limited opportunities to connect them to these managed taxonomic frameworks, increasing the likelihood that they will become out-of-date due to changing taxonomies almost as soon as they are created and deployed.

A worst-case situation for some identification resources, particularly born-digital ones, is that they become so out-of-date that they go offline and cease to be either managed or deployed. This results in the loss of a significant amount of work and investment.

In addition to identification resources being widely dispersed and disconnected, the taxonomic expertise that underpins them is also dispersed and declining, particularly in

[16] http://www.anbg.gov.au/photo/
[17] https://biodiversity.org.au/nsl/services
[18] https://biodiversity.org.au/afd/home
[19] http://www.environment.gov.au/science/abrs

some taxonomic groups[20]. This creates an urgent need to both capture existing diagnostic knowledge and maintain deployed resources when their original developers are no longer active.

The Vision

The vision of this implementation plan is to develop, within the decade 2018–2027, *a comprehensive, flexible, integrated, accessible service for identification of Australian organisms*, for a wide range of users.

Comprehensive means that the service will be capable of managing and deploying all available identification tools across all diagnostic modes and will be scalable to all Australian species.

Integrated means that the service will comprise a single portal, a one-stop-shop for identifications, and will provide seamless links between identification modes.

It also means that the service will be directly connected to other taxonomic services such as the agreed taxonomic backbone managed by ABRS, and the distributional data provided by the Atlas of Living Australia (ALA)[21], Australasian Virtual Herbarium (AVH)[22], Online Zoological Collections of Australian Museums (OZCAM)[23] etc.

Directly connecting the identification service to the agreed taxonomic backbone managed by ABRS will bring substantial efficiencies: for example, when a species name changes in the underlying taxonomic backbone, the name change will immediately flow through to the identification resources, rather than causing these to become out-of-date.

An integrated system optimally comprises interoperable modules or functions, each of which can be managed independently and used independently or in concert as appropriate.

This vision is for a set of modules that will map to the different identification modes discussed above, as follows:

Gestalt identification. Using the taxonomic backbone, a user will be able to enter the identification resource at any taxonomic level (if the specimen to be identified is clearly a beetle, the identification can start with all beetles as its taxon scope).

Visual pattern-matching. This module will comprise diagnostic images of taxa. At any stage in an identification, the set of images of all taxa in scope can be displayed.

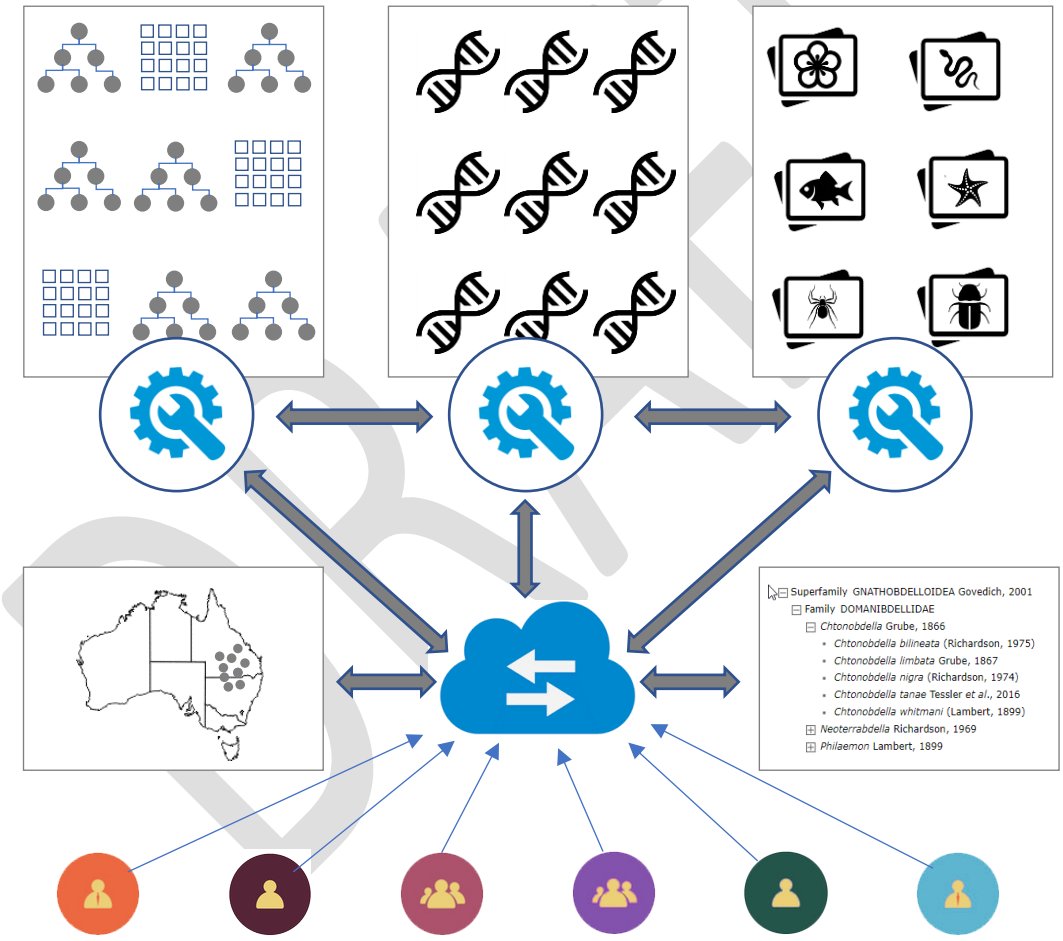
Trait-matching. Tree-based and matrix-based trait-matching keys will be combined into a single trait-matching module. Recent advances in developing a data model that encompasses both these identification modes will allow this integration.

[20] The decadal plan's Strategic Actions 6.1 and 6.2 address this issue, and will be the subject of another implementation plan.
[21] https://www.ala.org.au/
[22] https://avh.chah.org.au/
[23] http://www.ozcam.org.au/

Genome matching. A genome-matching module will interpolate a processing layer between the service and the GenBank and BOLD databases. This layer will store fitness-for-use annotations for records in GenBank and BOLD, providing a level of control over records for identifications in Australia that is not available in the underlying databases.

Machine learning. Machine learning will be the most difficult module to integrate, partly because the technology is developing rapidly and partly because it does not lend itself to comprehensive, taxonomically broad identifications. A machine learning module will be incorporated as a test environment to explore integration of this technology.

Geographic scoping: Connections to existing geography-based services including the ALA, AVH and OZCAM will be used to enable geographic scoping for the other identification modules.



The proposed integrated diagnostics capability. Identification resources (e.g. trait-matching keys, DNA and image-based matching resources) are managed together in modules. Web services exposed by each module provide interoperability, allowing an identification to be passed from one module to another as appropriate. Users interact with the system through an integration module, which also allows for geographic scoping and taxonomic management of the resources.

The integration of these modules with each other and with other services will allow an identification to begin in one mode, then be passed to other modes as appropriate. For example, an identification of a beetle found in an export grain shipment could proceed as follows:

1. Recognising that the specimen is a beetle (gestalt identification), the identification commences with a scope comprising all Australian beetles.
2. A gene sequence derived from the specimen is BLASTED against GenBank and BOLD (genome matching), which return moderate matches to six species spread across three genera within family A and a high match to a single species in family B.
3. The names of the two families are passed to a trait-based matching system, which provides a set of traits that differ between the two families. Based on these traits, the identifier determines that the specimen does not belong to family B and matches family A closely.
4. The trait-based matching system then provides a key to genera within family A. Using this, the identifier determines that the beetle belongs in a genus that is not represented in either GenBank or BOLD. The genus has three known species in Australia.
5. Images of representative specimens of these species are available in the system. Based on these, the identifier matches the specimen to one species (pattern matching). The species is not a biosecurity risk.
6. The specimen is vouchered, and its sequence uploaded to GenBank; the sequence of the species from family B is annotated as a dubious identification. An image of the specimen is uploaded to the system's image bank.

Similarly, an identification of a plant specimen collected by a citizen scientist in central Queensland could proceed as follows:

1. Knowing where the specimen was collected, a call is made to the ALA, which returns a checklist of all plants likely to occur in central Queensland (geographic scoping).
2. An image of the specimen along with the checklist is passed to a machine learning module, which excludes 90% of the species in the checklist, but fails to discriminate between the remaining 10% (machine learning).
3. The shortlist of candidate taxa is passed to the trait-matching module, using which the shortlist is reduced to three potential species (trait matching).
4. Images of representative specimens of these three species are available in the system. Based on these, the identifier matches the specimen to two species (pattern matching).
5. The image is sent to an expert in that taxonomic group, who determines that it belongs to one of the two species (gestalt matching).

Importantly, under this vision identification resources (traits, images etc), once assembled, will be permanently available and manageable. The problem of resources becoming either out-of-date and unmanageable, or going offline and unavailable, will be eliminated.

The system as envisioned will be scalable to all Australian organisms, and capable of integrating all legacy and future identification resources.

DRAFT

Tasks

The vision outlined here can be built in the following steps (see also Appendix 1):

1. A new program, the Australian National Biodiversity Diagnostics (ANBD) program, be established to coordinate implementation of this plan. This would be most appropriately established as a multi-agency program involving, at a minimum, the Commonwealth Department of Agriculture and Water Resources (DAWR), Plant Health Australia (PHA), the Australian Biological Resources Study (ABRS), the Atlas of Living Australia (ALA), and BioPlatforms Australia (BPA). Dedicated funding will be needed for this program (see Resourcing below).
2. A trait-based matching module be developed in concert with ALA and the development of a trait library for Australian taxa.
3. A campaign and program be developed for strategic capture of diagnostic images of Australian taxa. The campaign should focus on the capture of *diagnostic* images, not just any or all images. This implies strong quality control including taxonomic curation of images, and the capture of images that show diagnostic features. The campaign should have two phases:
    a. An aggregation phase, to index and make available all available diagnostic images from existing repositories. At the end of this phase, it will be possible for the first time to determine which Australian taxa have been imaged, and which have not.
    b. An extension phase, to capture images of taxa that have never been photographed (or are not diagnosable from existing images). This should engage citizen scientists working with and supported by taxonomic experts. There are many keen citizen scientists with digital cameras who would very likely engage enthusiastically with such a campaign, if suitable guidance can be given.
4. A services layer be developed that will enable efficient and effective DNA-based identification of Australian organisms. This will comprise a layer between the NSL and AFD managed by ABRS and the DNA databases managed by Genbank, BOLD and BioPlatforms Australia (BPA). The layer will enable sequences to be tagged with fitness-for-purpose, linked with vouchered specimens, and managed effectively in the face of changing taxonomies. This layer should be developed by BPA or under contract to BPA.
5. A machine learning partnership be established to develop a module for implementing machine learning approaches to biodiversity diagnostics. Several universities in Australia have active machine learning research programs, and these have the expertise to develop this capability.
6. An overarching integration module be developed to integrate all these and to enable identifications to be passed between modules. The integration module should be built around the taxonomic backbone developed and managed by ABRS.

DRAFT

Outcomes

If this vision is implemented, anyone in Australia will be able to identify any Australian organism to an appropriate taxonomic level, for any purpose, using the most appropriate identification mode.

Users will include, but not be limited to:

- Scientists, identifying specimens for research
- Biosecurity diagnosticians, identifying organisms of biosecurity significance in exports and imports
- Agricultural consultants, farmers etc, identifying organisms that threaten or benefit agriculture
- Environmental consultants, identifying taxa for environmental impact assessment
- Conservation practitioners, identifying rare and threatened, or threatening, taxa for conservation assessments and research
- Students, identifying organisms to become familiar with taxonomic groups of interest
- Members of the public including citizen scientists, identifying organism for interest or as part of citizen science survey programs

The currently widely scattered, often inaccessible, *ad hoc* and often poorly maintained identification resources in Australia will be managed in a coordinated, efficient and effective way, made more widely available than is currently possible, and managed in a way that will keep them up-to-date in the face of changing taxonomies. Development of identification and diagnostic resources will be able to be strategic, to target areas of greatest need while efficiently building on existing resources when available, and to use the identification mode that is most efficient, effective and appropriate for the task at hand. The integration that is key to this vision will also enable the appropriate handling of sensitive identifications, such as of trade-sensitive or conservation listed species.

Dependencies and linkages

The decadal plan's Strategic Action 2.1, comprising this vision, is closely linked with a range of other strategic actions, as expected given the central role identification plays in taxonomy and biosystematics. The following strategic actions are relevant:

Strategic Action 4.2: "*By 2028 we will have unified, authoritative checklists of all named species and other taxa in Australia, native and naturalised*".

A coordinated and well-managed taxonomic backbone is critical for realising this vision. Currently, two separate databases manage the names of Australian taxa – the NSL for plants, algae and fungi and AFD for animals. This strategic action will see these merged into a single service covering all biota. This will be maximally efficient for a wide range of uses including an integrated diagnostics capability.

Strategic action 4.3: *"We will build a curated, vouchered reference library of DNA sequences covering the breadth of the tree of life in our region".*

Such a curated reference library will be a key resource for implementing this vision. The reference library envisaged under SA 4.3 will comprise both Australian sequence databases and the international GenBank and BOLD databases as described above. The services layer described in this implementation plan will be a key component of the implementation of SA 4.3

***Strategic action 4.4**: "We will establish a freely accessible, authoritative, curated online image bank of the best available diagnostic images of Australian and New Zealand organisms".*

This strategic action closely matches the image campaign and program discussed above.

***Strategic action 4.6**: "By 2028 we will build a curated and well-managed trait library capable of capturing key ecological and morphological traits".*

This strategic action closely matches the trait-based identification module discussed above. Note however that a trait library that is capable of capturing ecological and morphological traits will not necessarily be optimal for contributing those traits to an identification system. The trait library will need to be developed with this important use case in mind.

***Strategic action 4.7**: "By 2028 we will have databased all botanical specimens, and at least half of all zoological specimens, in Australian and New Zealand biodiversity collections".*

Effective geographic scoping for identification requires a sound knowledge of the distributions of Australian organisms, as evidenced by occurrence records (vouchered and unvouchered) available in online resources.

Occurrence records for plants and many vertebrate groups in Australia (especially birds), made available through the ALA, AVH and OZCAM, have very good coverage and enable an accurate understanding of species distributions. However, for many other groups, particularly invertebrates, most available records are not online. Museum collections together hold more than 60 million of the 70 million specimens in Australian biodiversity collections, but only 5 million of these (8%) are available online through OZCAM. Until a significant databasing effort brings most of these records online, our understanding of the distributions of many species, and hence the ability to use distribution to help with identifications, is limited.

Note that even with good representation of specimen and observational records in databases, knowledge of the true distributions of species will always be limited by sparse or relatively sparse sampling. Species distribution modelling can ameliorate this problem and improve knowledge of the likely distributions of species. A collaborative project between Griffith University and the ALA is seeking to build an enhanced species distribution modelling capability for Australian taxa using ALA records. This may provide a substantial improvement in geographic scoping for diagnosis.

DRAFT

***Strategic action 6.1**: "We will engage with organisations to improve succession planning, mentoring and enhanced capabilities for the taxonomy and biosystematics sector in Australia and New Zealand".*

Ensuring that there is adequate taxonomic expertise in Australia to support biodiversity identifications is critical for several reasons. Firstly, well-trained taxonomists are required to build, curate and manage the identification resources dealt with here. Second, trained taxonomists will always provide a vital identification resource in their own right. This strategic action seeks to ensure that the next generation of taxonomists is adequately trained and fostered to prevent a decline in our diagnostic capabilities over time.

Risks

There are few technological impediments to realising this vision. Most of the resources needed for an integrated, comprehensive identification capability in Australia – a well-managed taxonomic backbone, good knowledge of distributions and traits, mature systems for deploying existing identification keys, abundant images in accessible repositories etc. – already exist for at least some taxonomic groups. The key tasks are to integrate these resources, make them interoperable, and work towards making them comprehensive.

A key risk therefore is that activities in the diagnostics space continue to be *ad hoc*, disconnected and non-strategic.

This risk will be exacerbated if resourcing for strategic development continues to be extremely limited and piecemeal. It can be ameliorated by establishing a dedicated program with sufficient funding to provide leadership and strategy and to coordinate activities across the multiple existing agencies (Commonwealth and State and Territory) that develop and use identification services.

The most severe risk for the vision described here is failure to secure adequate, ongoing and sustainable resourcing. This can be ameliorated by working to recruit key champions in the biosecurity, taxonomy, agriculture and environmental policy and planning sectors, by establishing a defined program to coordinate advocacy for the vision, and by ensuring up-front recognition that long-term sustainable funding (for strategic coordination, management and curation of diagnostics resources) as well as short-term program funding (for development of specific, targeted resources and systems) are required.

Responsibilities

A key requirement to bring this vision to fruition is the establishment of a dedicated program to champion it and coordinate the multiple agencies and modules required for true diagnostic integration.

The proposed Australian National Biodiversity Diagnostics (ANBD) program would have responsibility for:

- raising development funding to create the interoperable identification modules that comprise the system

DRAFT

- coordinating a program of work to bring existing identification resources into the interoperable system
- identifying gaps in diagnostic capability and development of a strategy for the creation of new diagnostic tools to fill these gaps
- coordinating the development of new content (identification keys, targeted sequencing, imagery) to fill these gaps, and
- coordinating ongoing curation of content to ensure that it remains fit-for-purpose in the face of new taxonomic knowledge

Key partners in the ANBD would have responsibilities as follows:

DAWR: coordination to ensure that identification resources critical for biosecurity are prioritised, and to integrate and develop identification resources relevant to biosecurity and agriculture.

ABRS: coordination to ensure that key identification resources across the breadth of Australian biodiversity are strategically developed for conservation, science and policy needs, and management and maintenance of the taxonomic backbone that underpins the diagnostics system.

ALA: integration of the ANBD with other Atlas data resources including traits, descriptions, images and observational records, and coordination of an imagery module and program of work to integrate existing image repositories and coordinate a citizen science campaign for targeted imaging.

BPA: coordination of a metadata layer module to ensure that DNA sequences in GenBank and other online repositories are tagged with fitness-for-purpose and quality-control tags, and coordination with ALA of development of an operational sequence-based identification module.

Other partners in this program would include:

Council of Heads of Australasian Herbaria (CHAH) and Council of Heads of Australian Faunal Collections (CHAFC): coordination of ongoing and targeted taxonomic research and of development of strategic identification resources across all modules.

Plant Health Australia (PHA): advocacy and coordination of this diagnostics program in the context of biosecurity.

Taxonomy Australia: Advocacy and coordination of this diagnostics program in the context of the taxonomy decadal plan.

State and CSIRO herbaria and museums: ongoing taxonomic research, development and curation of content and diagnostics resources, and curation with ABRS of a national architecture of taxa and their names.

DRAFT

Synergies with other programs

The development of identification and diagnostic resources is a key activity across a wide range of government-funded programs including biosecurity, agricultural extension and research, conservation planning, and taxonomy and biosystematics.

Currently, diagnostic activity in many of these sectors is *ad hoc* and poorly coordinated, resulting in identification resources that are widely scattered and not or scarcely interoperable.

The enhanced, integrated capability described here has potential to revolutionise the way in which identification resources in Australia are created, deployed, managed and maintained. For this reason, if the system envisaged is built it is expected that business-as-usual activities across many of these sectors will transition to the new system, spreading the funding and work burden widely while at the same time bringing efficiencies (both of scale and resulting from the integration) to these activities.

Many projects to build identification resources for Australian organisms exist or have been proposed, and many others will be commenced over time. The short-term nature of most of these means that listing current projects is outside the scope of this implementation plan; however, these are important for the success of this vision. Small-scale projects such as the development of specific identification keys, imaging programs and DNA sequencing projects, will play a significant role in creating content, while larger-scale projects[24] could potentially take responsibility for building or extending some modules.

One area with a more coordinated approach to the development and deployment of diagnostic resources is the biosecurity sector. The intersection of existing capabilities with the enhanced capability envisaged in this plan is outlined in Appendix 2.

Timeline

Development of an improved, integrated diagnostics capability and platform for Australia is urgent, to support critical needs particularly in biosecurity and conservation and to ensure that diagnostic knowledge and other content is captured from an ageing and declining workforce of Australian taxonomists before key skills are lost.

Fortunately, Australia has high capability for developing the integrated system envisaged here. The architecture proposed (of independent but interoperable modules for each of the primary diagnostics modes) lends itself to parallel development. With sufficient resourcing, the proposed system could be developed within a 2-year development timeframe.

[24] A current proposal to build an identification key for all plants of the Australian savanna, led by James Cook University, is large enough that it would require development that overlaps substantially with the system proposed here. Similarly, the Flora of Australia project (www.ausflora.org.au) built as a collaboration between the ALA, ABRS and Council of Heads of Australasian Herbaria, provides a clear use case for an integrated identification capability and has strong synergies with this project

Development and curation of content (traits, images, sequences etc) is an ongoing requirement and is not included in this timeline.

Timelines for individual components are as follows:

Trait-based matching module. A proof-of-concept has been developed for this module as a desktop java application. This needs to be developed further and stress-tested to ensure that it is scaleable to the needs of this proposal (early testing has indicated that it is highly scaleable). Assuming it passes this test, an online version needs to be developed. Java libraries developed for the proof-of-concept will be re-useable for this.

Estimated time for development of an online implementation: 12 months, 1 developer

DNA sequence repository metadata layer

As discussed, a key limitation to the use of existing DNA sequence repositories, particularly GenBank, is variable (sometimes poor) taxonomic quality control and limited linkage of sequences to vouchered specimens. This module will store fitness-for-use annotations for records in GenBank and BOLD, providing a level of control over records for identifications in Australia that is not available in the underlying databases. A combination of GenBank's APIs[25] and this fitness-for-use layer will facilitate the use of DNA sequence matching for accurate identification of Australian organisms.

Because this will be a new module, detailed specifications will be needed before commencing, to establish standards and benchmarks for achievable and useful annotations. This will need to be completed before work begins on the development of the annotations layer and its interface.

Estimated time for development of specification and annotation standards: 6 months Estimated time for development of layer and interface: 12 months, 1 developer

DNA sequence identification module

While DNA sequences are routinely used for identification of organisms in some specialised cases, there few good examples of fully operationalised, online DNA sequence identification tools. Most are bespoke, lab-based protocols that cannot be readily generalised or opened for broad use.

The module envisaged here will enable anyone to submit a sequence in a suitable format through a web form, then use a BLAST search against GenBank, filtered using the metadata

[25] https://www.ncbi.nlm.nih.gov/home/develop/api/

layer discussed above, to provide the best possible identification to the lowest reliable taxonomic level.

Again, detailed specifications will be needed before commencing this module. These, and the module development work, can proceed in parallel with development of the annotation layer discussed above.

Estimated time for development of specification: 6 months Estimated time for development of module and interface: 12 months, 1 developer

Images module

Three key problems limiting the use of images for identification are (1) unquantifiable gaps in available images (that is, some taxa even in well-known groups such as plants have never been photographed, but it's not currently possible to readily tell which these are), (2) existing images in available repositories vary widely in diagnostic quality, and (3) existing image repositories have very simple search functions and do not allow the assembly of images against a set of taxa (e.g. the set of taxa in scope for a particular identification step).

This module needs to address all three of these problems. Most effective will be the assembly of high-quality diagnostic images against a taxonomic backbone: if this is done it will be possible to tell which taxa in the backbone have no images adequate for diagnosis (problems 1 & 2), and the taxonomic backbone will enable ready assembly of available images for any taxonomic scope.

Development of this module is relatively simple: it requires a hierarchy (graph) of taxa (available from the ABRS taxonomic backbone) and a database of images or image references annotated for diagnostic quality and referenced to the taxon hierarchy.

Estimated time for specification and development of module: 12 months, 1 developer

Overarching integration module

Integration of the diagnostics modules outlined above is relatively straightforward if close attention is paid to interoperability when the modules are developed (that is, if each module exposes an API through web services, with suitable standards to enable interoperability).

If each module can output a list of the taxa (controlled against the underlying taxonomic backbone) that are in scope at any step in an identification, and each module can accept a list of taxa as the taxonomic scope for an identification step, then an identification can be readily passed from one module to another.

Hence, the integration envisaged in this plan only requires an overarching wrapper interface that will allow users to invoke any module, and pass an identification between modules.

DRAFT

Estimated time for specification and development of module: 12 months, 1 developer

Resourcing

Some aspects of a diagnostics capability for Australia can be achieved without dedicated resourcing; however, the integrated, comprehensive capability described here will need resourcing to build and maintain.

Critically, investment of carefully targeted and planned resources now will bring substantial cost savings over time, compared with continuing with the current *ad hoc* approach to diagnostics in Australia. Savings will be achieved through the efficiencies an integrated system brings for creating, maintaining and using the diagnostic tools, and from reducing the risk or pre-empting potential outbreaks of pests, loss of species, or loss of commodities markets from inadequate biosecurity and pest and environmental management.

Four separate components of resourcing need to be addressed:

1. Managing the Australian National Biodiversity Diagnostics Program
2. Building the interoperable identification modules
3. Bringing existing (legacy) identification resources into the system, to allow effective curation and to make full use of the advantages of interoperability, and
4. Ongoing taxonomic curation in the face of new taxonomic knowledge (e.g. new species, changes to existing taxonomic concepts)

1. Managing the Australian National Biodiversity Diagnostics (ANBD) Program

The ANBD would coordinate development funding, content creation and long-term strategic management and maintenance of the enhanced diagnostics capability described here.

It should be established initially with 3 FTE, comprising a Director, support officer and systems analyst. The office of the ANBD would be responsible for budgeting and contracting development and content work, systems analysis and assessment to ensure that individual modules are interoperable, coordination of standards development as required, and strategic planning for content (identification resource) development.

2. Building the interoperable system

As outlined under Timelines above, if an Australian National Biodiversity Diagnostics program is established for facilitation and management, an integrated system of interoperable modules to support a comprehensive diagnostics capability for Australia could be built in an estimated 18 months of development time using 5 FTE developers and a systems architect.

DRAFT

3. Managing legacy identification resources

There are many existing identification resources (keys, images, sequences etc) widely scattered in multiple formats. It is not feasible to budget to bring all these into the interoperable system envisaged here.

However, if the system as developed includes an effective pipeline for importing existing identification resources, and enough resources are brought into the system to demonstrate its value, then it is likely that other parties will see a self-interest in using the system and will manage their content into it. This happened with the development of the KeyBase platform for handling dichotomous keys – the platform was developed and populated with one set of keys, and is now used by multiple partners and agencies for managing their own key sets.

A workplan based on 2 FTE content coordinators for 2 years is estimated to be sufficient to bring enough existing resources into the interoperable system to demonstrate its value and build support for its wider adoption across a broad spectrum of content curators.

4. Ongoing content curation

The advantages of the system described here are expected to result in wide uptake by creators and curators of identification tools in Australia. As discussed above, this has already been seen with the KeyBase system, which is substantially less flexible and maintainable then the system described here. For this reason, owners of identification tools who use the system are likely to maintain their content as core business to a large extent.

However, ongoing curation of taxonomic content such as identification resources is required, because taxonomy is an ongoing activity and taxonomies change over time. While some content will be self-managed by third parties (individual taxonomists, biosecurity diagnosticians, institutions etc), other content will become orphaned as a result of retirements of taxonomists or closure of new projects. Curation of content is as important as creation of new content, but is rarely considered when budgeting projects.

For this reason, a key part of the Australian National Biodiversity Diagnostics Program should be ongoing support for content curators. Support needed will depend on the uptake and rate of growth of the system, and is likely to grow over time. For this reason, it cannot be adequately budgeted at this time, but needs to be considered in planning.

DRAFT

Summary of resourcing

The Australian National Biodiversity Diagnostics program could be established, a flexible, integrated, accessible service for identification of Australian organisms could be built, and a significant start made on building a diagnostics capacity that is scalable to all Australian organisms, with the following nominal budget[26]:

	Year 1	Year 2	Year 3 and ongoing
ANBD Director	1 FTE	1 FTE	1 FTE
ANBD Support Officer	1 FTE	1 FTE	1 FTE
ANBD Systems Analyst	1 FTE	1 FTE	
Developers	4 FTE	1.5 FTE	0.5 FTE
Content coordinators	2 FTE	2 FTE	?
Total	9 FTE	5.5 FTE	2.5+ FTE

Summary

A comprehensive, flexible, integrated, accessible service for identification of Australian organisms does not currently exist but is feasible and could be built with a modest investment.

Importantly, Australia has very high capacity to build such a service, which would be the world's best identification resource. The Australian taxonomy and biosystematics community has a long history in the development of world-leading and innovative identification tools. The Australian biosecurity system is world-class and provides a strong use-case for effective, efficient and timely diagnosis and identification. The decadal plan for taxonomy and biosystematics, which provides the overarching framework for this implementation plan, establishes a strong context for the program of work described here.

[26] Note that resourcing estimates here are expressed as FTE rather than dollar figures. Suitable position levels for host agencies will need to be established to develop an itemised budget for this proposal.

DRAFT

Appendix 1. Tasks and actions

Task 1: *Establish the Australian National Biodiversity Diagnostics (ANBD) program to coordinate implementation of this plan.*			
Description	Diagnostics and identification in Australia are currently developed, deployed and maintained in an *ad hoc* manner, and this constitutes a substantial limitation on current capability. The ANBD would reverse this. Its proposed main roles are to: • raise development funding to create the interoperable identification modules that comprise the system • coordinate a program of work to bring existing identification resources into the interoperable system • identify gaps in diagnostic capability and develop strategies for the creation of new diagnostic tools to fill these gaps • coordinate the strategic development of new content (identification keys, targeted sequencing, imagery) to fill these gaps, and • coordinate ongoing curation of content to ensure that it remains fit-for-purpose in the face of new taxonomic knowledge		
Outcomes	Short-, medium- and long-term strategic development, management and maintenance of the enhanced diagnostics capability described here.		
Actions		Potential Lead	Priority	Duration/ Timeframe
1.1 Establish a Steering Group to build support among partners for the ANBD		DAWR; ABRS; PHA	Very High	6 months
1.2 Develop funding proposals for Budget round 2020/21		Steering Group	Very high	12 months
1.3 Establish the ANBD		Steering Group	Very high	18 months
1.4 Manage the ANBD		Steering Group	High	Ongoing
Potential Partners				
ALA, CSIRO, SPHD, PHA, State and Territory herbaria and museums, Taxonomy Australia				
Total estimated resources: $480,000				
$160,000 per annum for 3 years to appoint and maintain a National Coordinator and support activities of the Steering Group				

DRAFT

Task 2: *Develop an integrated trait-based matching module*			
Description	Trait-based matching (tree- and matrix-based keys) are and will remain important diagnostic resources, especially if interoperable with other identification modes as described here. Currently, however, trait-based keys are not well-managed or maintained. A more sophisticated platform for their deployment and maintenance is needed.		
Outcomes	A flexible, efficient and interoperable platform for managing, deploying and maintaining trait-based identification keys		
Actions	Potential Lead	Priority	Duration/ Timeframe
1.1 Develop and test a scalable, integrated and interoperable trait-based identification module	Taxonomy Australia	High	12 months
1.2 Progressively upload legacy trait-based identification tools into the new module	ABRS	High	2 years
1.3 Manage, maintain and expand content in the module	ABRS, State and Territory herbaria and museums	High	ongoing
Potential Partners			
DAWR, ABRS, State and Territory herbaria and museums			
Total estimated resources: $330,000			
$110,000 p.a. for 1 year to develop the module and $110.000 p.a. for two years to upload legacy keys			

DRAFT

Task 3: *Initiate a campaign for strategic capture of diagnostic images of Australian taxa*			
Description	Diagnostic images provide key confirmatory resources for diagnosis, as well as providing a useful identification mode in their own right in some cases. However, as with other identification modes, diagnostic images are currently *ad hoc*, widely scattered and poorly managed. Because of this, many images are not diagnostic, and there are currently no mechanisms for determining which taxa have images and which do not, and which available images are diagnostic.		
Outcomes	A well-managed and interoperable online library of diagnostic images, managed in the context of the taxonomic hierarchy of organisms in Australia and scalable to the Australian biota		
Actions		Potential Lead	Priority	Duration/ Timeframe
1.1 Establish and develop guidelines and quality standards for diagnostic images		ANBD, ALA	Medium	6 months
1.2 Enhance the Atlas of Living Australia image repository and develop a front end for attaching diagnostic images to taxa in the taxonomic backbone		ALA	Medium	12 months
1.3 Aggregate and index existing images from other repositories		ALA	Medium	2 years
1.4 Engage citizen scientists to work with taxonomic experts to capture diagnostic images for taxa that are not adequately imaged using existing resources.		ANBD, ABRS	Medium	ongoing
Potential Partners				
DAWR, Australian National Botanic Gardens (manages the Australian Plant Image Index), State and Territory herbaria and museums, Taxonomy Australia				
Total estimated resources: $330,000				
$110,000 p.a. for 1 year to develop the module and 110,000 p.a. for 2 years to aggregate content and engage citizen scientists				

DRAFT

Task 4: *Develop an annotations layer and identification module that will enable efficient and effective DNA-based identification of Australian organisms*			
Description	While many DNA sequences for Australian organisms are available on international repositories such as GenBank, their use for identification and diagnosis is in many cases very limited due to the poor taxonomic curation of these databases and historically inadequate standards for vouchering of samples. A services layer that allows sequences to be tagged for fitness-for-purpose, linked with vouchered specimens, and managed effectively in the face of changing taxonomies, would provide a way to mitigate this problem.		
Outcomes	A services layer that tags sequences in GenBank and BOLD with fitness-for-purpose annotations and taxonomic determinations, increasing machine (and human) discoverability and use of these sequences for diagnostics.		
Actions		Potential Lead	Priority	Duration/ Timeframe
1.1 Develop specifications and annotation standards		ANBD	High	6 months
1.2 Develop the services layer to enable annotation of sequences with fitness-for-use tags for Australian biodiversity diagnostics		BioPlatforms Australia	High	12 months
1.3 Develop a module for enhanced DNA diagnostics using the services layer and BLAST searching.		BioPlatforms Australia, ALA	High	12 months
1.4 Tag sequences		ANBD	Medium	Ongoing
Potential Partners				
ALA, State and Territory museums and herbaria, Taxonomy Australia				
Total estimated resources: $220,000				
$110,000 per annum for 2 years to develop the modules (assumes funding for ANBD)				

DRAFT

Task 5: *Assess and develop machine learning approaches to biodiversity diagnostics*			
Description	Machine learning has substantial potential in diagnostics, as evident by mobile apps and online applications used for identification in some use cases. However, the limits of machine learning in this space have not been adequately tested, and no large-scale machine-learning applications have been developed in Australia.		
Outcomes	An assessment of machine learning for broad-scale application in biodiversity diagnostics and identification, including the limits (if any) to the technology		
Actions	Potential Lead	Priority	Duration/ Timeframe
1.1 Build a partnership with university machine-learning researchers to assess potentials and limits of machine learning in real-world identification applications	ANBD	Medium	2 years
Potential Partners			
Universities in Australia and overseas			
Total estimated resources: uncertain			
Assumes funding for ANBD			

DRAFT

Task 6: *Develop an overarching integration module to allow interoperability between diagnostics modules*			
Description	Integration and interoperability both within and between identification modules are key to this proposal. The integration module needs to include a taxonomic backbone, to ensure that all identification tools share a common taxonomy, and standards to enable an identification task to be passed between modules.		
Outcomes	The integration and operability that are critical to providing the step-change in diagnostics and identification capability in Australia developed in this proposal.		
Actions	Potential Lead	Priority	Duration/ Timeframe
1.1 Establish and implement standards to ensure interoperability between modules	ANBD	High	6 months
1.2 Ensure integration of all modules with the accepted taxonomic backbone	ABRS, DAWR	High	6 months
Potential Partners			
Taxonomy Australia			
Total estimated resources e.g. $110,000			
$110,000 per annum for 1 year to develop the module			

DRAFT

Appendix 2. Intersection with Australian biosecurity diagnostics capabilities.

This implementation plan describes an enhanced biodiversity identification and diagnostics capability that covers the breadth of the Australian biota and is relevant to a wide range of use cases. One important use case is biosecurity and agricultural diagnostics – the identification of organisms that are a real or potential threat to agriculture, the environment, and animal, plant and human health.

Given its breadth of scope, the implementation plan touches on but does not focus specifically on biosecurity diagnostics. Given the importance of the biosecurity use case, this Appendix outlines intersections between this implementation plan and existing biosecurity diagnostics capabilities, agencies and partnerships.

Key intersections are as follows:

1. The *Agricultural Competitiveness White Paper*[27] provides an overarching policy, strategy and (to June 2019) funding framework for biosecurity diagnostics in Australia. The White Paper provided $200 million for biosecurity surveillance and analysis, mostly allocated over 4 years to 30 June 2019 with some information and analysis components continuing to 30 June 2020.

Under the theme *Growing scientific capability*, the White Paper improved plant and animal health diagnostics capability and infrastructure by:

- upgrading laboratory infrastructure
- coordinating and sharing diagnostic information
- identifying capability gaps for plant and animal pests and disease diagnosis
- providing diagnostic training
- developing diagnosis tools.

This enhanced capability is clearly complementary to, and enhances the use case for, the integrated biodiversity diagnostics and identification service described in this implementation plan.

2. The *Department of Agriculture and Water Resources* (DAWR), including the offices of the Australian Chief Plant Protection Officer (ACPPO)[28] and Australian Chief Environmental Biosecurity Officer (ACEBO)[29], is a key federal department dealing with biosecurity and biodiversity as it relates to agriculture.

There are key synergies between biodiversity in agriculture and biodiversity in the non-agricultural environment – each affects the other, and there is a strong need for biodiversity diagnostics and identification in both arenas. Maintaining separate diagnostics frameworks for biosecurity and agriculture on one hand, and biodiversity and taxonomy on the other,

[27] https://agwhitepaper.agriculture.gov.au/
[28] http://www.agriculture.gov.au/plant/health/acppo
[29] http://www.agriculture.gov.au/biosecurity/environmental/cebo

works against these synergies. Working with the Department of the Environment and Energy (particularly the Australian Biological Resources Study), with CSIRO, and with State and Territory museums, herbaria, and agriculture departments, DAWR can play a lead role in championing, developing and managing the biodiversity diagnostics capability described in this implementation plan.

2. *Plant Health Australia*[30] is a not-for-profit company established in 2000 to coordinate a government-industry partnership for plant biosecurity in Australia. PHA facilitates a strong government and industry biosecurity partnership to minimise pest impacts on Australia, enhance market access and contribute to industry and community sustainability.

PHA manages the Australian Plant Pest Database (APPD)[31], which aggregates voucher specimen and other information from 18 contributing pest databases. With its experience managing APPD, PHA is an important partner to develop and manage some components of the system described here

3. The *Plant Health Committee* (PHC)[32] of DAWR is the peak government plant biosecurity policy and decision-making forum, and has responsibility for delivering on national priority reform areas including those identified for implementation of the Intergovernmental Agreement on Biosecurity (IGAB) and overseeing the implementation by governments of the National Plant Biosecurity Strategy (NPBS). PHC has a clear role, working in partnership with the biodiversity community (museums and herbarium collections institutions), in setting the framework for the Australian National Biodiversity Diagnostics program described here, and in championing and representing its benefits to government.

3. The *Subcommittee on Plant Health Diagnostics* (SPHD)[33] is a subcommittee of the Plant Health Committee (PHC) and includes representation from the Australian, state and territory governments, Plant Health Australia, CSIRO and the New Zealand Ministry of Primary Industries. Among other roles, SPHD coordinates the development of National Diagnostic Protocols for priority plant pests, coordinates the National Plant Biosecurity Diagnostic Network (NPBDN)[34] and assists the development of diagnostic tools and material. SPHD has a key role in helping coordinate platform and content development for the Australian National Biodiversity Diagnostics program.

SPHD also took on responsibility in 2017 for maintaining PaDIL[35] – an online portal of images and information on pests – pending the development of a business plan to determine its future role and viability. PaDIL is an important repository and a useful test case for the enhanced library of diagnostic images which comprises one module of the integrated system described here.

[30] http://www.planthealthaustralia.com.au/
[31] http://www.planthealthaustralia.com.au/resources/australian-plant-pest-database/
[32] http://www.agriculture.gov.au/plant/health/committees/phc/
[33] http://www.plantbiosecuritydiagnostics.net.au/work/subcommittee-on-plant-health-diagnostics/
[34] http://www.plantbiosecuritydiagnostics.net.au/
[35] http://www.padil.gov.au/