Supplementary information 1: Criteria used for species selection

The criteria itemised below were used to make the choice of the species featuring in the case study.

Criterion	Additional explanation
Species represents an alien invasive species in Australia and elsewhere	Species could have ecosystem and/or other environmental impacts, and/or has spread widely outside of its native range. Focus on Australia because Atlas of Living Australia is the national biodiversity database used as one of the two test case infrastructures in the present study.
Species is of compelling scientific interest	With relevance to management, policy and research.
Range of taxonomic groups	Spread species selection across plants, insects, birds, etc.
Sufficient data exist, and are in public repositories	Sufficient data, both spatially and temporally needed to make the investigation valid. Data must be accessible.
Data in different places, and different data in those different places	Necessary to challenge the exploration of workflow and infrastructures issues.

Supplementary information 2: Subspecies of *Acacia longifolia*

Taxonomic name variations, synonyms and naming errors were reviewed and standardised, as was available information on the native geographic range of each species. These steps often constitute a significant hurdle in collating and synthesising data on invasive alien species (McGeoch *et al* 2012). For example, *Acacia longifolia*, currently comprises two subspecies (*A. longifolia* subsp. *longifolia* and *A. longifolia* subsp. *sophorae (Labill.) Court*) and includes historic records under different and separate species names (Butcher *et al* 2001). The Sydney golden wattle (*A. longifolia* subsp. *longifolia*) is very similar to the coastal wattle (*A, longifolia* subsp. *sophorae*) and they intergrade naturally (Butcher *et al* 2001). Both subtaxa have naturalised distributions beyond their native range, with *A. longifolia* subsp. *longifolia* more widely distributed beyond its native range and *A. longifolia* subsp. *sophorae* generally restricted to near coastal habitats (DEEDI 2016a, 2016b). Therefore, both subtaxa were retained for analysis and *A. longifolia* without a subspecies epithet were assumed to be subsp. *longifolia*. 'Cultivated' occurrences of these taxa were removed where noted in the source data. An additional challenge is lack of resolution on the historic boundaries of the native geographic ranges of these taxa (McGeoch *et al* 2012). Known sites of introduction of the subspecies were putatively designated as 'alien', and core regions of the native range as 'native'. Occurrence records situated close to the known, coarse scale boundaries of the subspecies or in regions where the native/introduced status were unclear. *A. longifolia* subsp. *longifolia* has become so widely naturalised across its native range in southern and eastern parts of Australia that broad determinations of native or alien using distribution records are difficult to make. On the other hand, the narrower niche and native distribution described for *A. longifolia* subsp. *sophorae* was more readily designated. The alien ranges of the two subspecies were combined for the analysis. Informative descriptions of the original and naturalised distributions of the two taxa are provided by the 'Weeds of Australia' online fact sheets (sources A and B below). The Australian Native Plant Society provides a range map for *A. longifolia* subsp. *Sophorae* (source C below). These resources guided the range designations. A generalised native range map for *A. longifolia* can be found on Wikipedia (source D below), assumed indicative of *A. longifolia* subsp. *longifolia*. Comprehensive determination of native and invasive population occurrences would require field assessment.

Sources:

(A) https://keyserver.lucidcentral.org/weeds/data/media/Html/acacia_longifolia_subsp._longifolia.htm

(B) https://keyserver.lucidcentral.org/weeds/data/media/Html/acacia_longifolia_subsp._sophorae.htm

(C) http://anpsa.org.au/a-sop.html

(D) https://en.wikipedia.org/wiki/Acacia_longifolia

Butcher P, Chapman A, Conn B, Court A, Cowan R, George A, Hill R, Keith D, Kodela P, Leach G, Lewington M, Mcdonald M, Macphail M, Maslin B, Pedley L, Ross J, Tame T, Tindale M and Wilson A 2001 Flora of Australia Volume 11B: Mimosaceae Acacia Part 2 (Melbourne, Victoria, AU.: CSIRO Publishing)

DEEDI 2016a Weeds of Australia - Biosecurity Queensland Edition Fact Sheet: Acacia longifolia subsp. longifolia *Dep. Employment, Econ. Dev. Innov. (DEEDI), Identic Pty Ltd., Biosecurity Queensl.* Online: https://keyserver.lucidcentral.org/weeds/data/media/Html/acacia_longifolia_subsp._longifolia.htm

DEEDI 2016b Weeds of Australia - Biosecurity Queensland Edition Fact Sheet: Acacia longifolia subsp. sophorae *Dep. Employment, Econ. Dev. Innov. (DEEDI), Identic Pty Ltd., Biosecurity Queensl.* Online: https://keyserver.lucidcentral.org/weeds/data/media/Html/acacia_longifolia_subsp._sophorae.htm

McGeoch M A, Spear D, Kleynhans E J and Marais E 2012 Uncertainty in invasive alien species listing Ecol. Appl. 22 959–71 Online: http://doi.wiley.com/10.1890/11-1252.1

Supplementary information 3: Detailed results from executing IVSD

Detailed results obtained from executing the IVSD workflow are tabulated in Tables 1 – 3 below. The step numbers in column 1 of the tables correspond to the step numbers in Figures 3 – 5 of the article main text. Any notes mentioned in tables can be found at the end of each table.

Specific data files, and other related artefacts created by steps of the work have been grouped together by step into a data object package. This package is available from the Zenodo repository with the doi: <doi reference to where it can be found, once deposited>.

Table 1. **Results of first stage of IVSD workflow for selected three species, making data EBV-usable.**

Step		Sydney golden wattle (*Acacia longifolia*)	European wasp (*Vespula germanica*)	Cattle egret (*Bubulcus ibis*)
1	Data license check	Data from ALA and GBIF is open licensed, using Creative Commons licenses.		
2	Locate access	Principal entry points to ALA and GBIF data are, respectively the websites www.ala.org.au and www.gbif.org.		
3	Select initial taxa of interest	ALA returns 19,432 occurrences with 36 name variants. GBIF returns 1,251,090 occurrences with 66 name variants.	ALA, returns 542 occurrences with 1 name variant. GBIF, returns 5,337 occurrences with 1 name variants.	ALA returns 105,561 occurrences with 3 name variants. GBIF returns 995,580 occurrences with 3 name variants.
4	Resolve taxonomic name issues	*Acacia longifolia* has two subspecies *A. l. longifolia* and *A. l. sophorae*. Synonym were resolved easily. *Acacia longifolia* without subspecies were assumed *A. l. longifolia*.	Names resolve unambiguously.	ALA and GBIF use different naming conventions, *Ardea ibis* and *Bubulcus ibis* that resolve unambiguously.
5	Delimit occurrence data by internal filtering (NOTE)	GBIF: 12,849 records. ALA: 7,729 records.	GBIF: 3,247 records. ALA: 413 records. Observations in South Africa filtered out due to data quality issues.	GBIF: 730,554 records. ALA: 48,660 records.
6	Export data	Data exported as a CSV file and downloaded to local computing resources for further processing.		
7	Remove unnecessary columns (ALA data only)	Using MS Excel, removed 148 columns with null or no useful information.	Using MS Excel, removed 133 columns with null or no useful information.	Using MS Excel, removed 151 columns with null or no useful information.
8	Repair field names, values and map to standard vocabularies	The field names provided by ALA and GBIF were cleaned and mapped to the standardised Darwin Core and Dublin Core vocabularies and "NULL" was replaced with the empty string.	The field names provided by ALA and GBIF were cleaned and mapped to the standardised Darwin Core and Dublin Core vocabularies and "NULL" was replaced with the empty string.	The field names provided by ALA and GBIF were cleaned and mapped to the standardised Darwin Core and Dublin Core vocabularies and "NULL" was replaced with the empty string.
9	Save data	Data is usable for producing EBV data products. It could be preserved with a log of actions performed (its provenance) and published with a persistent identifier such as a digital object identifier (DOI).		

NOTE: Using the GBIF portal, the only filter applied at step 5 was a taxonomic filter to ensure return of all records interpreted as representing the species (including records published under known synonyms for the species, based on the GBIF taxonomy). Further filtering and validation of GBIF data for geography, time-period and other aspects was deferred until later in the workflow. Using the ALA portal however, multiple filters were applied. Records were removed if any of the following criteria were met: country not Australia; scientific name did not include the subspecies; the location was generalized; coordinate precision was out of range; scientific name not in national checklists; the location was in the ocean or a river; if the record was a duplicate; had an incomplete or invalid collection date; coordinate uncertainty was out of range; the nominated State/Territory did not match the coordinates; an unrecognized geodetic datum; the occurrence status (present, absent) unrecognized; or the country nominated did not match the provided coordinates.

Table 2. **Results of second stage of IVSD workflow for selected three species, making data EBV-ready.**

Step		Sydney golden wattle (*Acacia longifolia*)	European wasp (*Vespula germanica*)	Cattle egret (*Bubulcus ibis*)
10	Create Darwin Core Archives	For each file, the best primary key field, dwc:catalogNumber, was located and a Darwin Core Archive was created based on its index. (ALA coreId: 11, GBIF coreid: 5)	For each file, the best primary key field, dwc:catalogNumber, was located and a Darwin Core Archive was created based on its index. (ALA coreId: 26, GBIF coreid: 290)	For each file, the best primary key field, dwc:catalogNumber, was located and a Darwin Core Archive was created based on its index. (ALA coreId: 26, GBIF coreid: 290)
11	Merge Darwin Core Archives	The Darwin Core Archives were merged based on matching the values for the coreId fields to an archive containing 20,577 records.	The Darwin Core Archives were merged based on matching the values for the coreId fields to an archive containing Total: 3,660 records.	The Darwin Core Archives were merged based on matching the values for the coreId fields to an archive containing Total: 779,214 records.
12	Identify standardised populated fields	379 mapped fields, of which 200 have at least one record with values.	594 mapped fields, of which 222 have at least one record with values.	684 mapped fields, of which 289 have at least one record with values.
13	Filtering on valid geo-reference and date	20,080 records are geo-referenced with valid year, split into two subtaxa: subsp. *longifolia* 9,956 records and subsp. *sophorae* 10,124 records.	3,458 records are geo-referenced with valid year.	777,214 records are geo-referenced with valid year.
14	Profile data (spatial)	There are 14,302 distinct latitude and longitude locations referenced. There are 8,568 distinct locations for subsp. *longifolia* and 6,095 for subsp. *sophorae*.	There are 2,182 distinct latitude and longitude locations referenced.	There are 147,213distinct latitude and longitude locations referenced.

Step		Sydney golden wattle (*Acacia longifolia*)	European wasp (*Vespula germanica*)	Cattle egret (*Bubulcus ibis*)
15	Profile data (temporal)	There are 157 years represented in the data, between 1770 and 2017. *A.* subsp. *longifolia* is recorded in 139 years (1770-2017) and *A.* subsp. *sophorae* is recorded in 140 years (1837-2017).	There are 109 years represented in the data, between 1879 and 2017.	There are 103 years represented in the data, between 1770 and 2017.
16	Retain useful fields	Retain 39 column fields as potentially useful.	Retain 46 column fields as potentially useful.	Retain 54 column fields as potentially useful.
17	Remove duplicate records, if appropriate (NOTE)	18,517 distinct records with year and spatial reference of initially 20,080, split into two subtaxa: subsp. *longifolia* 9,234 records and subsp. *sophorae* 9,283 records	2,689 distinct records with year and spatial reference of initially 3,458.	405,153 distinct records with year and spatial reference of initially 777,214.
18	Filtering based on assertions	Required valid location reference and year, and within 2000m of land – conservatively based on case study parameter using 2km grid for AOO (tested using GADM in GIS).		
19	Summary metrics	Occurrences of subsp. *longifolia* were recorded in 139 years between 1770 and 2017. Not all years recorded an occurrence; and occurrences in any year varied from 1 to more than 500. Occurrences of subsp. *sophorae* were recorded in 140 years between 1837 and 2017, though not in all years; and occurrences in any year varied from 1 to nearly 1,300.	Occurrences of *V. germanica* were recorded in 109 years between 1879 and 2017. Not all years recorded an occurrence; and occurrences in any year varied from 1 to 227.	Occurrences of *B. ibis* were recorded in 103 years between 1770 and 2017. Not all years recorded an occurrence; and occurrences in any year varied from 1 to more than 70,000.
20	Publish EBV-ready data product	Data is ready for producing EBV data products. It should be preserved with a log of actions performed (its provenance) and published with a persistent identifier such as a digital object identifier (DOI).		

NOTE: We didn't specifically apply any de-duplication for two reasons: i) it wasn't needed given the nature of the use case and focus on AOO by year; and ii) the inconsistent date format fields and corruption of those fields at some stage precluded accurate de-duplication. Therefore, we simply applied generated summary or aggregation (resolution) based on locations (latitude, longitude) within years and counted individual records as occurrences at that temporal resolution. Here we report the summary results.

539 **Table 3. Results of third stage of IVSD workflow for selected three species, specific to intended**
540 **application; in this case to provide invasion trends of three priority invasive species.**

Step		Sydney golden wattle (*Acacia longifolia*)	German / European wasp (*Vespula germanica*)	Cattle egret (*Bubulcus ibis*)
		Retain 17 column fields potentially relevant and limit dates to year of observation (Step18_ExtractData_Acacia_longifolia.xls and Step18_ExtractData_Acacia_sophorae.xls)	Retain 15 column fields potentially relevant and limit dates to year of observation (Step18_ExtractData_Vespula_germanica)	Retain 17 column fields potentially relevant and limit dates to year of observation (Step18_ExtractData_Bubulcus_ibis.xls)
21	Distinguish native and alien ranges	View data in GIS, with supporting information land and administration boundaries (e.g., GADM) and published descriptions / maps of native ranges, add field and annotate native / alien range.		
22	Generate time-series look up tables	Create look up tables for relating year to decade and other time periods for summary analysis (e.g., quarter and half century) and join with taxon data.		
23	Calculate AOO by decade	Use projected and gridded land data to assign Area of Occupancy (AOO) to taxon records per year and decade. (GLOBISB_AOO_calcs_figures_V2.xls)		
24	Summarise	Produce maps and graphs of results.		

541

542

543

Supplementary information 4: Additional issues encountered

In addition to principal difficulties mentioned in the article text, the following is a complete list of difficulties encountered as the workflow was manually executed:

1. Both GBIF and ALA as publishers of data have terms of use agreements that can include more specific data provider terms covering the data they supply. Checking the licensing conditions (step 1) for the data retrieved from GBIF and ALA proved challenging by manual means to ensure specific terms were precisely followed at the record level, for example in respect of mandated attribution of use.
2. Many workflow steps required expert human judgement in combination with computer assistance for manipulation, making them tedious and error-prone. Different procedures had to be applied according to whether the step was performed on ALA or GBIF data and in the ALA or GBIF environment.
3. Selecting occurrence records, geographical boundaries and time periods of interest (steps 3 – 10) had to be executed in the environment of the data publisher to delineate a first data set of relevance. Removing blank / irrelevant fields and checking fitness for use of individual data records could only be applied with third-party tools (e.g., spreadsheet software) once data had been retrieved from the publisher's repository.
4. The extent to which selection and filtering of available data were carried out *in situ* within the environment of the data publisher, versus how much of that was done after all potentially relevant data has been retrieved is a matter of search strategy and capability of the publisher's environment. Insufficient attention to this at the start led, initially to incompatible sets of records from both sources, GBIF and ALA. Those from GBIF included spatially invalid records, which had subsequently to be filtered out whereas those from ALA did not. The same occurrence records are rendered differently in GBIF and the ALA. The same fields may be named differently in different agencies and many record fields are not even exposed in the interfaces making comprehensive filtering impossible.
5. Because all record-level filtering could not be achieved the GBIF or ALA website environments, the data had to be exported as a CSV (comma-separated variable) file and a third-party tool such as a spreadsheet or database tool used to remove unwanted columns or filter on fields not exposed in the web interfaces.
6. Steps 14 and 15 (Profile data) at present involve specialised third-party tools but these could also be part of the offered service. Before finally downloading data, these tools should be available as sanity checks, with an opportunity to add extra filters. This would mean that steps 10 (apply spatial and temporal filters) and 11 (data quality checks) should also be fully handled within the infrastructure.
7. Use and insertion of supplementary information, such as on alien / native status of a given occurrence could be improved progressively, based both assessment of the different purposes for which data is demanded and as integration of multiple biodiversity information resources matures over time.
8. Merging records from both publishers (steps 10 – 12) required bespoke computer programming scripts to be written. Careful attention had to be given to detect overlaps and gaps between GBIF and ALA data, due to lack of alignment of column headers between GBIF and ALA exported data.
9. The merging process was further complicated by the lack of a single standard field for uniquely identifying records. The main candidate for matching records would most likely be dwc:occurrenceID. However, many data providers instead put their unique identifier in the dwc:catalogNumber or dwc:eventID fields. In cases where the data provider does use dwc:occurrenceID, there was still an issue. ALA records that have been published to GBIF have their occurrenceID values replaced with an internal ALA universally unique identifier value, but ALA downloads have the original dwc:occurrenceID value.
10. In merging, there were limited cases where matches were found to indicate duplicate records.
11. Some issues with date references not being standardised in ISO 8601 format or being corrupted were encountered; although in many cases a valid year reference was available in parsed fields allowing use of the records in the case study.
12. When reviewing data fields in the merged CSV files for relevance to the case study (steps 19 – 23), it was found that the supplementary fields differed for each taxon; in part related to differences in community of practice data collection standards applied in the use of Darwin Core terms.

593 13. The users were ‘overwhelmed’ by the confusing array of fields from the merger of the GBIF and ALA
594 data, meaning further field selections had to be made and checked for usability in the analysis. Without
595 sophisticated user-interface tools to explore and segment the data, for example linked tabular and spatial
596 faceting, it was too hard.

597 14. Absence of information in the primary biodiversity data about the native / alien status of a species as
598 observed, and thus the need to consult third-party sources such as the CABI Invasive Species Compendium
599 highlights that building EBV data products depends on multiple data sources of different kinds.

600 15. Accurate recording of work done (i.e., provenance) to a level of detail enough to easily replicate the work
601 was manual and therefore difficult and time consuming.

602
603

604

605

606 # Supplementary information 5: Time-series occupancy maps

607 An archive package (zip file) provides individual maps for the three taxa in 'PNG' image format (600dpi) with
608 content summarised with captions and attribution in power point slides. – To be added.