

Appendices: Data fitness for use in research on alien and invasive species

Final task group report

McGeoch MA, Groom QJ, Pagad S, Petrosyan V, Ruiz G & Wilson J (2016) Appendices to *Data fitness for use in research on alien and invasive species*. Copenhagen: GBIF Secretariat. Available online at: <u>http://www.gbif.org/resource/82958</u>.

Appendix A: GBIF survey of data use for research on alien and invasive species and the results generated from the survey

Introduction

As part of a broader global strategy to support applications of biodiversity data, GBIF has convened a Task Group to assess data fitness for use in research on alien and invasive species. GBIF is an international open data infrastructure, funded by governments.

In this survey, use of the terms alien and invasive are broad, and included the multiple related terms used in the field of invasion biology, such as introduced and non-native. This call was to complete the survey to improve the suitability and access to data for broad scale use in invasion biology and the fields related to it (such as ecological and social research contexts). The survey was designed by the Task Group to capture the experience of data users and publishers aiming to document limitations in existing GBIF services, improve the utility of GBIF-mediated data, and suggest improvements in the functionality of GBIF for specific needs.

This survey is an appendix to the report from the Task Group on Data Fitness for Use in Research on Invasive Alien Species

Methods

Each of the task group members identified 10 or more "power users" of the GBIF mediated data in the A&IS context. The next wave of the responses, starting from early September, was through GBIF e-communication channels, NEOBIOTA and IUCN congress 2016.

The survey was made as short as possible, expected to be under about 20 minutes. Survey logic was implemented preventing a subset of users from responding to a subset of the questions in the case where they have not had any experience with using GBIF-mediated data. A number of survey questions was made obligatory.

Data from the questionnaire was superficially analysed in R (version 3.3.2) using the package *psych* (Version 1.6.8; Revelle, 2014). Ordinal variables were coded as numbers to make them analysable. Correlation analysis was used to identify the related variables and the *iclust* function was used to implement a simple hierarchical cluster analysis. To make the analysis comprehensible the demographic data was compared against the questions related

to current features and uses of GBIF, separately to the questions related to future needs of users.

Results

Survey results are summarized below in Figures A1–A21 and Tables A1–A3, followed by the summaries of the free text answers, and are discussed in the main report.





Weekly (Starting on the date)





Figure A2. Paired correlations of question related to improvements to GBIF. A chart showing the paired correlations between the different answers to the question on improvements to GBIF. On the diagonal are histograms of the answers. On average these histograms suggest all the improvements are important, but the native ranges stand out.



Correlation plot

Figure A3. A colour visualization of the paired correlations of the question related to improvements to GBIF. Correlations are plotted in colour to improve comprehension. Notable features include the negative correlation between researchers who have published and those interested in management data; the positive correlation between those interested in impact, habitat and mechanism of introduction and the positive correlation between those that have published and use the API.

ICLUST



Figure A4. A hierarchical cluster analysis of responses. A simple hierarchical cluster analysis of responses to the question on improvements to GBIF. This uses the R package iClust to cluster responses and characters of the respondents. Two classes of respondents can be seen. One group has published with GBIF-mediated data, use the API and use GBIF regularly. The other group is much more interested in the ecological information and impact of invasive species.

Correlation plot



Figure A5. A visualization of the pair correlations for the questions related to the features of GBIF and the purpose for which people use GBIF. Note that researchers who have published with GBIF mobilized data are negatively correlated with features such as taxonomy and country pages.





Figure A6. A hierarchical cluster analysis of responses to the question on features of GBIF and the purposes for which GBIF is used. A simple hierarchical cluster analysis of responses to the questions on the purpose people use GBIF for and the features they use. This uses the R package iClust to cluster responses and characters of the respondents. Three classes of respondents can be seen. One group has published with GBIF-mediated data, uses the API and use GBIF regularly. Another group browses the species pages and taxonomy, while the third group is interested in the impact of invasive species and uses GBIF for risk assessment.



Figure A7. Q2: Do you consider your organization commercial? Answered: 218. Skipped: 0.



Figure A8. Q3: What realms do you work in? Answered: 218. Skipped: 0.



Figure A9. Q4: What is your disciplinary expertise? Answered: 218. Skipped: 0.



Figure A10. Q5: Role. Answered: 218. Skipped: 0.



Figure A11. Q6: What species groups do you work with? Answered: 218. Skipped: 0.



Figure A12. Q7: How long have you been using GBIF resources? Answered: 141. Skipped: 77.



Figure A13. Q8: How regularly do you use GBIF? Answered: 141. Skipped: 77.



Figure A14. Q9: How do you access GBIF-mediated data relevant to alien and invasive species? Answered: 119. Skipped: 99.



Figure A15. Q10: For what purpose have you used GBIF-mediated data, and how useful were the data? Answered: 141. Skipped: 77. Axis X – Weighted average based on the five step scale of usefulness (extremely useful to not useful at all).

	Extremely useful	Very useful	Somewhat useful	Slightly useful	Not at all useful	N/A	Total	Weighted Average
Taxonomy	9.93% 14	20.57% 29	20.57% 29	7.80% 11	2.13% 3	39.01% 55	141	2.53
Extracting lists of species	14.18% 20	27.66% 39	13.48% 19	6.38% 9	0.71 % 1	37.59% 53	141	2.23
Species distributions	24.82% 35	29.79% 42	24.82% 35	5.67% 8	0.00% 0	14.89% 21	141	2.13
Modelling	12.77% 18	18.44% 26	14.18% 20	4.96% 7	2.84 % 4	46.81% 66	141	2.37
Georeferencing	9.93% 14	15.60% 22	22.70% 32	7.80% 11	0.71 % 1	43.26% 61	141	2.54
Gap analysis	4.96% 7	12.77% 18	14.89% 21	4.26% 6	0.00% 0	63.12% 89	141	2.50
Searching for sources of data	10.64% 15	24.11% 34	24.82% 35	1.42% 2	2.13% 3	36.88% 52	141	2.37
Invasion horizon scanning	8.51% 12	12.06% 17	18.44% 26	7.80% 11	1.42% 2	51.77% 73	141	2.62
Risk assessment	8.51% 12	19.86% 28	21.28% 30	7.09% 10	1.42% 2	41.84% 59	141	2.54
Trends in abundance or distribution	10.64% 15	13.48% 19	19.86% 28	8.51% 12	2.13% 3	45.39% 64	141	2.60

Table A1. Q10: For what purpose have you used GBIF-mediated data, and how useful were the data?



Figure A16. Q11: Which features of GBIF (GBIF.org) are useful for your work on biological invasions? Answered: 137. Skipped: 81. Axis X – Weighted average based the five step scale of usefulness (extremely useful to not useful at all).

	Extremely useful	Very useful	Somewhat useful	Slightly useful	Not at all useful	N/A	Total	Weighted Average
Maps	27.27%	31.82%	16.67%	8.33%	1.52%	14.39%		
	36	42	22	11	2	19	132	2.12
Web	15.57%	23.77%	15.57%	5.74%	0.82%	38.52%		
services	19	29	19	7	1	47	122	2.23
Taxonomy	18.40%	20.00%	16.80%	12.80%	0.80%	31.20%		
	23	25	21	16	1	39	125	2.38
Country	12.90%	20.16%	20.97%	12.90%	1.61%	31.45%		
pages	16	25	26	16	2	39	124	2.56
Dataset	15.32%	25.81%	21.77%	7.26%	0.81%	29.03%		
pages	19	32	27	9	1	36	124	2.33
Species	22.48%	34.88%	14.73%	6.98%	0.78%	20.16%		
pages	29	45	19	9	1	26	129	2.11
Record	15.70%	28.93%	19.01%	9.09%	0.83%	26.45%		
pages	19	35	23	11	1	32	121	2.33
Verbatim,	7.50%	11.67%	20.83%	8.33%	3.33%	48.33%		
uninterpreted	9	14	25	10	4	58	120	2.77
data								
Multimedia	9.24%	11.76%	11.76%	11.76%	5.88%	49.58 %		
(images,	11	14	14	14	7	59	119	2.87
sounas, video)								
/								

Table A2. Q11: Which features of GBIF (GBIF.org) are useful for your work on biological invasions?





	Extremely useful	Very useful	Somewhat useful	Slightly useful	Not at all useful	Total	Weighted Average
Habitat	33.09% 46	41.01% 57	21.58% 30	2.16% 3	2.16 % 3	139	1.99
Species traits	31.85 % 43	44.44% 60	11.85% 16	8.89% 12	2.96 % 4	135	2.07
Mechanism of introduction	38.81% 52	32.84 % 44	17.91% 24	8.21 % 11	2.24% 3	134	2.02
Native ranges	50.37% 68	35.56% 48	8.89% 12	3.70% 5	1.48% 2	135	1.70
Absence data	37.40% 49	33.59% 44	16.79% 22	6.11% 8	6.11% 8	131	2.10
Abundance	38.93% 51	37.40% 49	16.79% 22	5.34% 7	1.53% 2	131	1.93
Filling geographic gaps	49.61% 64	31.01% 40	13.95% 18	4.65% 6	0.78 % 1	129	1.76
Filling taxonomic gaps	29.01% 38	32.82% 43	25.95% 34	7.63% 10	4.58% 6	131	2.26
Impact information	36.57% 49	35.07% 47	14.93% 20	8.21 % 11	5.22% 7	134	2.10
Management information	27.41% 37	31.11% 42	25.93% 35	11.11% 15	4.44% 6	135	2.34

Table A3. Q12: In which of the following areas would data improvements be most useful for research on biological invasions?



Figure A18. Q13: Which other sources of information on biological invasions do you use?



Figure A19. Q15: Are you aware of sources that would improve data coverage of alien and invasive species in GBIF? Answered: 141, Skipped: 77.



Figure A20. Q17: What spatial resolution of data is most useful in the following categories? Answered: 102. Skipped: 116.



Figure A21. Q18: Have you published any papers/reports that reference GBIF information or data? Answered: 115. Skipped: 103.

Open ended comments from respondents

The following section summarises the five main topics (A–E) and comments from question 12 and 20 of the survey (N=54). The numbers in brackets {} reflects the total number of responses on this topic, split by question in some cases. The number of asterisks (*) next to each statement indicate the number of time the same or very similar comment was made. @,#,\$ – points that are closely related to each other

Summary of Q20 and Q12. Additional recommendations on how GBIF can become more useful to alien and invasive species research and applications; In which of the following areas would data improvements be most useful for research on biological invasions?

A. Strategy {2}

There is a range of information highly relevant to A&IS that are currently not accommodated by GBIF. The extent to which GBIF should be expanded to accommodate these variables, expanded to provide links to alternative sources that provide such information, or decide that certain information is beyond the scope of GBIF forms part of the recommendations of this report.

- A strategic recommendation is to render the list of tasks feasible by focusing efforts on a subset of taxa considered to be a high priority, i.e. high priority A&IS.
- Metadata (traits, pathways and impacts) should not be main focus
- Adopt strategic, staged approach to improving fitness for use by initial focus on selected high priority taxa.

B. Improve data coverage, currency, completeness and quality {23 +4 }

The most common themes were the need for better data coverage and data quality control. Content and process refinements identified included definitions, A&IS data standards, scale and resolution of data and speed/frequency of record uploads.

- Improve geographic coverage, fill gaps in data coverage, more georeferenced records, focus on poorly covered areas, distribution data is a GBIF priority *******
- Expand taxonomic coverage, e.g. plant pathogens, diseases
- Actively source existing literature collations of occurrence data; target, encourage and enable specific contributions of A&IS data *
- More frequent national updates, enable rapid publication of new records
- Provide more links to reference and related information sources
- Information on validation of scale of data based on source, resolution of georeferencing, spatial resolution of records **
- Clear, unambiguous definitions of invasion terms, also search terms ambiguous *
- Adopt data standards for A&IS *
- Accuracy: Ensure data are correct and thorough (quality control, taxonomic and geographic errors, duplicates, data verification) @ *****

C. Expanded/additional information content {28+7}

The most frequently identified information need, and most frequent comment overall, is information on species native ranges and, as follows, introduced range. This comment is reinforced by the call for information that forms the foundation of knowledge on species native and non-native ranges, i.e. absence data, data of introduction, eradication records and species range dynamics.

There were also calls for a number of types of information related to knowledge of species impact, including abundance, invasion status and priority, legislative status by country and management options. Finally, respondents requested a range of additional life history information such as habitat use, physiology and other species traits.

- More detailed explanation of taxonomic uncertainties
- Functional traits, physiology, habitat associations, cultivars
- Species interactions (host associations, biocontrol agents)
- Original collector information
- Date of introduction
- Absence data (also distinguish absence of data from absence of species) ******
- Include survey/sampling effort estimates*
- Capture range dynamics # *
- Distinguish current from historical distribution, flag alien/native status of records, info on native ranges, NB data on native range (native, non-native or unknown)\$
- Capture eradications

- Abundance data*
- Impact data **
- Invasion status/priority of species/risk assessment\$ *****
- Legislative status of species in country
- Management approaches

D. Expanded or refined functionality {10 + 3}

Suggested expanded functionality in order of frequency includes additional and more flexible data filters, ease and fitness for use of data extraction by filter. Expanded functionality on error descriptions and reporting, and the possibility of a surveillance or rapid response/reporting tool were suggested (see comments on range dynamics above).

- Critical annotation of data records, e.g. flag outliers, corrected records and those from species in captivity/botanical gardens @
- Enable users to report errors
- Enable filters by selected taxa/invasive species (and filter exclusions, e.g. fossils), countries and data sources; enable output of invasive species as textual lists ****
- Enable user defined geographic areas
- Make is easier to use, easier to extract and contribute information (Plugin for QGIS, tool to make it easier to construct API queries) **
- Allow users to map biogeographic regions over the data (e.g. MEOW ecoregions)
- Column formatting in data downloads
- Surveillance, rapid response tools#

E. Communication, collaboration and engagement {13+1}

The viewpoint that GBIF could do more to improve its visibility as a provider of A&IS data was fairly common, and that it should actively advertise and engage with a range of relevant potential partners, related activities and data publishers. The important role of GBIF to contribute to the formulation of data standards for A&IS was highlighted (see also below).

- Establish/flag a specific A&IS entry route (or specific page/website), into the Facility
 **
- Engage more with local information providers, link with country Focal Points, NGOs, expert networks and relevant journals; greater collaboration with other information providers/platform (e.g. USGS, NAS)*****
- Avoid duplication of effort (e.g. GISD, GRIIS, EOL)
- Promote international standards for A&IS information and work with other data publishers to achieve this
- Improve the visibility and credibility of the site for A&IS information, advertise GBIF *
- Incentivize data provers

Appendix B: Background and detailed rationale underpinning key topics and selected recommendations for improving data fitness for use on alien and invasive species

1. Proposed changes to Darwin Core for alien and invasive species (A&IS) research

What is Darwin Core?

Darwin Core (DwC) is a key data standard used for the dissemination of biodiversity observation data, most notably by the Global Biodiversity Information Facility (GBIF). As such, its structure acts as a filter through which data are passed and it can have a profound influence on the availability and quality of shared data. For this reason it is critical that DwC is made suitable for the data needs of invasion science and conservation.

The challenge

However, in practise the process of converting observation data to DwC can seem rather procrustean. Although DwC is a versatile standard, it is still in the process of evolution and lacks some finesse. A balance needs to be found between flexibility and standardization. On the one hand, flexibility helps data publishers find data fields to map their data, but on the other hand, too much freedom dilutes the usefulness of data and the benefits of standardization.

Data from all sorts of biodiversity projects and sources must be reshaped into the data fields and structures provided by DwC, optimally, without losing valuable information in the process. The potential data sources are vast and include data from the whole earth, every living organism and all habitats. Furthermore, these data are collected for various reasons and those reasons have an influence on the formats and types of data collected. Therefore, DwC has a difficult role to play in providing a standard suitable for all observation data. Providing data for research on alien and invasive species (A&IS) is only one of these roles, so improving DwC for A&IS research needs to be done in such a way as to not reduce its usefulness for other research areas and to encourage contribution from a wide variety of data holders, not just those interested in biological invasions.

Data needs in invasion science

There are many types of information that are needed for science and policy on A&IS and many of these are already covered within DwC, including specific details of taxonomy and location. However, there are three basic pieces of information that are regularly collected but lack adequate provision in DwC:

- 1. Whether the organism is native or alien to the location
- 2. Whether it currently exists at the location
- 3. How was it originally transported to this new location

Existing DwC vocabulary and suggested changes

establishmentMeanshttp://terms.tdwg.org/wiki/dwc:establishmentMeans

This term already exists within Darwin Core and is defined as "*The process by which the biological individual(s) represented in the Occurrence became established at the location.*" The suggested controlled vocabulary includes the terms "native", "introduced", "naturalised", "invasive" and "managed".

Our view is that this definition and the suggested vocabulary are mismatched and that the concepts of nativeness, management, invasion pathway and invasiveness should be in separate fields as they represent different concepts. We advocate adopting the invasion pathway categorisation terminology recommended by the <u>Convention on Biological Diversity</u> for the establishmentMeans field (see <u>https://www.cbd.int/doc/meetings/sbstta/sbstta-18-09-add1-en.pdf</u>). This will provide a richer terminology for data on introduction vector and a greater degree of standardization.

Furthermore, the need to make this change also highlights the concept of nativeness and the need for a dedicated data field for this within Darwin Core (see <u>origin</u>).

originhttp://s3.amazonaws.com/wiki_docs/Presence, Seasonal and Origin Attributes for Species Ranges.pdf

An expression of whether something is indigenous to an area is currently only expressed in DwC in the current controlled vocabulary of <u>establishmentMeans</u>. However, by separating the concepts of whether something is indigenous from how it came to be at the current location enables an important and clear distinction of the organism's origins at a site. Nativeness is considered an important criterion for the conservation of organisms when considering their importance to the biodiversity of a region.

We propose adding a new 'origin' field to DwC adopting the <u>origin</u> field of the <u>IUCN</u>. This simple vocabulary is already used by the conservation research community and aligning DwC and the biological invasions research community with this vocabulary has many advantages for standardization. There are however a few problems that need to be resolved or at least explicitly recognised.

The IUCN has the following terms for the field origin, "native", "reintroduced", "introduced", "vagrant" and "unknown". But in the context of biological invasions, it is not entirely clear what the term "vagrant" means. It could refer to a population that is introduced but not yet naturalised, but in a conservation context it can refer to individuals that arrive through natural dispersal, and as such are not introduced in the sense that there is no human action involved. The Vocabulary is therefore too simple to convey the range of situations relevant to invasion and conservation biology. However, there are very many significant advantages of adopting an already accepted 'standard' vocabulary that the limitations may be acceptable. The compromise is worth it to ensure interoperability, It will also only ever be a "suggested vocabulary", rather than a standard that is enforced.

One issue that remains to be resolved is whether there should also be a distinction between introductions that occurred before and after the modern era. Certainly in Europe this distinction is frequently made and has conservation value. In Europe the modern era is at the beginning of the Columbian exchange in 1500. For other parts of the world, such as New

Zealand, this date maybe later, and in other parts of the world the distinction is often much less clear, e.g. Asia and Africa.

occurrenceStatushttp:///h

This field already exists in DwC and has a suggested vocabulary of "present", "absent", "common", "irregular", "rare" and "doubtful". Though this information is useful it can equally well be expressed in DwC using the field individualCount or a combination of organismQuantity and organismQuantityType. However, in the particular case of species checklists there is no current means to express if something still exists in the area, or has become extinct. For this reason we suggest changing the controlled vocabulary of occurrenceStatus to the vocabulary used by the IUCN term 'presence'. This term has the simple vocabulary "extant", "possibly extinct", "extinct", "extinct post 1500" and "presence uncertain".

We also recommend that the DwC <u>documentation</u> is updated to make it clear that this term is intended to be used for species checklists, rather than observations. This clarification of the terms of this field is not required specifically for A&IS research, however, combined with the addition of the origin term and the establishmentMeans is does enable a much clearer expression of the status of native and alien organisms in a region or at a site.

Example

The <u>Alien Plants of Belgium website</u> contains a full list of alien plants for Belgium and details of their introduction and status. To publish this checklist on GBIF the fields need to be mapped to Darwin Core. Below (Table B1, B2) is a small section of the checklist and how it should be interpreted using the proposed vocabularies and new origin term (Verloove, 2016).

Taxon	Mode of introduction	First Record	Most recent record	Origin	Degree of naturalization	Means of introduction
Acanthus spinosus	Deliberate	2016	2016	E AF AS-Te	Casual	Hort.
Amaranthus clementii	Accidental	1939	1949	AUS	Casual	Wool
Amaranthus albus	Accidental	1857	Ν	NAM	Nat.	Grain, wool
Oenothera angustissima	Accidental	1860	1884	NAM	Ext.	Ore

Table B1. A section of the checklist of the Alien Plants of Belgium

Table B2. Darwin Core interpretation of the checklist illustrating the loss of relevant

 information and level of accuracy during translation from the checklist to DwC.

Taxon	dwc:origin	dwc:occurenceStatus	dwc:establishmentMeans
Acanthus spinosus	vagrant	extant	horticulture
Amaranthus clementii	vagrant	extinct	container/bulk
Amaranthus albus	introduced during the modern era	extant	seed contaminant container/bulk ¹
Oenothera angustissima	introduced during the modern era	extinct	container/bulk

1. Note that fields in Darwin Core that require multiple entries take a pipe () delimited string.

Other fields in Darwin Core perhaps relevant to alien species issues are listed below in Table B3.

Table B3. Darwin Core fields relevant to alien species issues.

MeasurementOrFact	Occurrence			
measurementID	organismQuantity			
measurementType	organismQuantityType			
measurementValue	lifeStage			
measurementAccuracy	reproductiveCondition			
measurementUnit	Behavior			
measurementDeterminedBy	establishmentMeans			
measurementDeterminedDate	occurrenceStatus			
measurementMethod	associatedTaxa			
measurementRemarks	occurrenceRemarks			
Organism	ResourceRelationship			
associatedOccurrences	resourceRelationshipID			
	resourceID			

2. Importance of Open Access Data to A&IS research

Research on A&IS is highly dependent on access to data from a wide variety of sources. Not only are data needed at the point of impact, but also from other counties. Furthermore, fine resolution data, both historical and contemporary, are needed. Needless to say no researcher could gather such data without the help and generosity of a wide range of stakeholders. In fact, such a large number of disparate actors collect these data that the only feasible way of sharing these data is to make them open access on a common repository (Groom et al. 2015).

For biodiversity observations this common repository is GBIF and to facilitate sharing they have implemented a system of data licensing that gives three choices to publishers. The first option is to make their data public domain and waive all rights to the data. A second option is to use a Creative Commons attribution licence

(<u>https://creativecommons.org/licenses/by/2.0/deed.en</u>), and the third option is to use the Creative Commons non-commercial attribution licence

(<u>https://creativecommons.org/licenses/by-nc/2.0/</u>). This system is a compromise between users who want simple access and those data publishers who want to ensure acknowledgement and to retain commercial rights to their data.

In the case of A&IS any restrictions placed on data makes them harder to use, and impedes the automation of workflows that could be used for impact assessment and early warning. This is particularly important for the large number of countries with poor data infrastructure and limited capacity and those that don't house their own A&IS data. Research on A&IS generally needs as much and as high quality data as possible. As a result, users will probably have to conform to the licensing requirements of the most restrictive publishers to ensure compliance. Policy makers should be aware that commercial organizations, including not-for-profit organizations, would be severely restricted in their ability to use all biodiversity observation data.

The following approaches/principles should be promoted: Wherever possible data publishers should put their data in the public domain. If this is not possible then publishers should consider time limits to the restriction of commercial usage rights. Data users should appropriately cite all data publishers, whether or not the licensing requires it. Funders should stipulate the data licensing required for funded projects.

3. Next generation sequencing and its importance to A&IS research

The application of next generation sequencing to the detection of organisms in environmental samples has enormous potential for the observation of biodiversity and a wide range of applications relevant to A&IS (Chown et al. 2015). These techniques are increasingly being applied, and it is anticipated that the growth in data volume will be exponential. This will provide new information on both well-known organisms and currently obscure taxa. In the realm of A&IS research, these techniques are likely to provide unprecedented monitoring, particularly of species once difficult to monitor, such as those in aquatic environments. Nevertheless, from a data management perspective, these data present many problems. Furthermore, these problems need to be addressed with some urgency to ensure that this young research field approaches data management robustly from the beginning.

Conventionally, GenBank is the repository for molecular sequence data, whereas GBIF is the repository of observation and specimen data. Given the potentially large number of observations from this source it is important these data are available to ecologist who would not normally search GenBank for observation data. Both the sequence, temporal, methodological and geographic information needs to be preserved to the same standards that we expect for other biodiversity observations. Ideally we should ensure that the results of eDNA research are preserved without losing information.

4. A&IS pathways and vectors

The focus on pathways of introduction aims to understand and prevent the transfer, invasion, and spread of A&IS. For management and policy, prioritization of pathways uses information on the full suite of vectors and routes by which alien propagules are introduced, and the propagule loads of such pathways (Carlton and Ruiz 2005), and is central to achieving Aichi Target 9 of the Strategic Plan for Biodiversity 2011–2020.

The mechanism of introduction for A&IS is a key data layer with diverse applications in research, management, and policy. Specifically, patterns of invasion are driven by particular modes of human-mediated dispersal, and these vary across taxonomic groups, geographic regions, and time periods. While research seeks to understand the resulting biogeographic patterns and ecological consequences, management and policy aims to reduce the extent and rate of species introductions. In each of these endeavours, identifying mechanism of introduction is a central piece of the puzzle.

We recommend that GBIF include data on mechanism of introduction, which is often referred to as "vector" or "pathway", and we use these synonymously here. As discussed in the main report text, this is a derived information field—based on the life history, time, and locations of species records—rather than an attribute of each occurrence record. As such, we recommend that GBIF develop key partnerships with data holders, who have been collecting these data. This could include checklists from individual countries, GBIF nodes, and data publishers (see Figure 2). Moreover, developing a mechanism for data exchange with GIASIP provides one efficient model for data exchange, regular updates, and use by a broad user community and the CBD. Moreover, this is feasible within a few years.

As noted above, we advocate adopting the invasion pathway categorisation terminology recommended by the Convention on Biological Diversity

(https://www.cbd.int/doc/meetings/sbstta/sbstta-20/information/sbstta-20-inf-05-en.pdf) for the establishmentMeans field (see also https://www.cbd.int/doc/meetings/sbstta/sbstta-18/official/sbstta-18-09-add1-en.pdf). A detailed categorization of pathways has been developed by the IUCN Invasive Species Specialist Group, and endorsed by the CBD (UNEP 2014; Essl et al. 2015). This hierarchical system encompasses three broad mechanisms and six principal pathway categories (Essl et al. 2015). Parties to the CBD have called for the use of this pathway framework for the purpose of assessing and prioritizing the risk posed by pathways (UNEP 2014), which will facilitate the reporting envisaged in Aichi Target 9 (McGeoch et al. 2016). It is particularly important to include the "SubCategory" level information for A&IS, because this provides the necessary granularity needed to understand invasion dynamics and evaluate performance of current management and policy actions, which operate at this level.

It is also critical to recognize that vector (pathway) does vary geographically and through time for the same species. For example, a species may be introduced in the ballast water of commercial ship to North America and live trade (e.g., seafood or bait export) to Australia or Asia. Further, an initial or primary introduction may occur to one bay by commercial ship, but secondary spread along a coastline or to islands may result from recreational boating or fishing. Thus, vector is not a species-level trait but specific to a species x location. As such, we recommend that the spatial scale of vector-level data be specified to hierarchical level, such as watershed, country, region, or continent, to refer to first recorded occurrence(s).

Finally, it is important to recognize multiple vectors are possible for introduction of a species to a location. This is seen in North America, where there is uncertainty about the vector responsible for initial introductions to the continent, such that ships' ballast water and hull biofouling (and sometimes additional vectors) are possible mechanisms for many species, based on life history and timing of arrival (Ruiz et al., 2011b, 2015). Thus, vector-level information must capture this real-world complexity to support fitness for use by the A&IS user community.

5. Alien-native range status

The most essential record-level attribute for A&IS is one that designates a record as falling within either the native or introduced range of the species (e.g. native range, introduced range, range history unknown). These designations may also eventually be associated with an estimate of certainty (GEO BON Species Populations Working Group). Date of record is also an important record-level attribute for A&IS, critical to capturing range dynamics and invasion trends.

Species alien status has been identified as one of the three Essential Variables for Invasion Monitoring, along with alien species occurrence and alien species impact (http://geobon.org/essential-biodiversity-variables/ebv-for-invasion-monitoring/). The rationale and priority of this variable or data type is provided in detail in Latombe et al. (2016), along with some general specifications for it collection and reporting, and therefore this information is not repeated here. However, within the Essential Biodiversity Variable framework (Pereira et al. 2013), species alien status is considered to be an attribute, or information ancillary to species occurrence—the two variables therefore go hand in hand. Both these variables are essential for a range of research and management applications, such as assessing the status of species along the introduction–naturalization–invasion continuum, delivering alert systems for new incursions, and model-based predictions of which species are candidates for future incursion (Latombe et al. 2016).

6. Annotation of A&IS using categorical information from other sources

There are a number of annotations, or 'attributes', associated with species occurrence records that are relevant and useful for research on A&IS (see for example Appendix B.4). These include 1) species-level attributes, 2) occurrence record-level attributes and also 3) species-level attributes filtered by a geographically defined attribute (geographic-level attributes).

 Species-level attributes include general attributes such as realm, functional group, habitat and host associations and other species-level traits. Species-level attributes of specific value to A&IS research include, for example, the presence/absence of the species on GRIIS; flagging a species as a biocontrol agent of a pest or weed; a list of biocontrol agents associated with an A&IS; a list of pathways of introduction and spread; standardised species impact categories.

As far as possible attribute data should use existing standards, methods or frameworks and their associated categories and lexicons, for example the Standard Categorisation of Pathways (UNEP 2014; Essl et al. 2015), Environmental Impact Classification for Alien Taxa (EICAT, Blackburn et al. 2014, Hawkins et al. 2015). In instances where GBIF is the primary

provider of a particular information attribute we recommend that an appropriate method/framework and lexicon is developed, published and adopted to establish its validity.

2. Geographic-level attributes: Some attributes may need to be expressed with reference to (or filtered via) geographic region associate attribute (e.g. country) rather than a species- or record-level attribute. For example, the alien/native status or regulatory status (e.g. 'nationally 'listed' species) of a species at country level. Survey data would also have geographic-level attribute data associated with it, such as survey metadata or species abundance estimates.

Several of these species-level and geographic-level attributes need to be supported by a systematic decision-making framework (see for example Ruiz et al. 1999, 2011a classifying and evaluating impacts by type and information quality) to deal with their inherent uncertainties and to demonstrate the evidence-base of the information and to support its transparency and repeatability as it is used and as it accumulates (McGeoch et al. 2012).

3. Record-level attributes: The most essential record-level attribute for A&IS is one that designates a record as falling within either the native or introduced range of the species (e.g. native range, introduced range, range history unknown). These designations may also eventually be associated with an estimate of certainty (GEO BON Species Populations Working Group). Date of record is also an important record-level attribute for A&IS (see Task 5 above).

Several habitat variables associated with occurrence records are especially valuable in characterizing, understanding, and predicting A&IS distributions and how these are shifting through time. The actual habitat distribution is unknown for a remarkable number of taxa; even where known, this also may change among geographic regions, as the degree of niche conservatism between native and introduced ranges is not clear.

We recommend several additional habitat and environmental attributes be collected routinely with occurrence records and included as part of data standards (see also Guralnick et al. 2016, in review). This should include standard terms for:

- Habitat type (soil, rock, water, air, vegetation type, etc.)
- Elevation (height / depth below surface or low tide)
- Simple environmental measures (air or water temperature, water salinity)

These same variables are of broad utility and are not specific to A&IS users. Thus, we recommend a process to evaluate, adopt, and promote inclusion of these core variables and standardization of nomenclature to meet the above objectives.

4. Some types of information will be relevant at all these levels. For example, it is useful to know which pathways a species has been transported along somewhere in the world (a species-level attribute), which of these pathways resulted in a species being introduced to any particular region (a geographic-level attribute), and the probable pathway and vector that resulted in any particular instance of occurrence (a record-level attribute). It is also useful to know which realm (terrestrial, freshwater, marine) and habitats a species occupies at both the species level and record level. The former can be invaluable to filtering data records, to flag outliers. The latter are invaluable to advancing and improving predictive ability, especially since habitat utilization is known to vary geographically.

7. Range dynamics captured using occurrence data

The presence records that constitute the occurrence of a species are commonly associated with a particular date of observation or record (which in some cases might be estimated, or might represent a 'latest' date of record). For A&IS, these records provide valuable information on how long a species has been present in a particular area (country or region) outside of its native range. Cumulatively, such occurrence records and associated dates provide valuable information on range expansion of particular A&IS, and trends in biological invasion across A&IS. Although this information is confounded by survey effort, investing in the capture of these data will become increasingly valuable as they accumulate and become more comprehensive. The research and infrastructure advancements required to support such information is being conducted, *inter alia*, under the GEO BON Species Populations Working Group.

Alien species occurrence has been identified as one of the three Essential Variables for Invasion Monitoring, along with alien status (status of a species or record as alien or native within a defined geographical unit or range) and alien species impact (Latombe et al. 2016). Alien species occurrence includes *taxonomically verified species presence or absence records at a locality with a geographic co-ordinate, or in a prescribed area, management or geopolitical unit or site* (Latombe et al. 2016). The rationale and priority of this variable or data type is provided in detail in Latombe et al. (2016), along with some general specifications for it collection and reporting, and therefore this information is not repeated here. However, this variable, i.e. alien species occurrence, is the **single most important variable** for delivery necessary to support monitoring and management and also the one that requires in-situ collection from countries both with and without capacity to do so. Such data are the core of GBIF's mandate, and GBIF is the most well positioned platform for housing open access information of this nature.

8. Species interactions relevant to A&IS

Species interactions profoundly influence the spread and impact of A&IS. The absence of obligate mutualists will prevent establishment, the presence of facultative mutualists can be an important stimulus for population growth rate and spread, and the presence of specific antagonists can limit and reduce an invasion. Currently GBIF does not hold such information in a way that is useful for A&IS research, nor are data publishers always providing these data. We see three main areas where improvements are needed.

First, there should be a data architecture that allows for observed interactions to be recorded (a record-level attribute). In most systems the organism interacted with is recorded in a free text notes field, with no standardisation of vocabulary or taxonomy. An exception is the popular online recording system iNaturalist (http://www.inaturalist.org). Within GBIF there are a couple of options to achieve a better structure for interactions data. Two organisms could be recorded similarly and an interaction term created on one or both to link between them. Darwin Core contains a field associatedOccurrences for linking two observations, however there is no mechanism to document the nature of the co-occurrence. Alternatively a record of a given species can contain a field for listing interacting organisms, i.e. the field associateTaxa in Darwin Core (http://terms.tdwg.org/wiki/dwc:associatedTaxa). Both

approaches will require an ontology for recording interactions (e.g. <u>http://www.ontobee.org/ontology/RO;</u> Smith et al. 2005).

Second it is important that there is a method to identify whether or not species co-occur at any given site (a geographic-level attribute). GBIF will then be able to provide important biotic context information that is linked to particular sites. There are a wide range of interaction data that could be useful (e.g. host-parasite relationships, competitive relationship, and quantitative food webs). In practice, however, it will be important to focus on a few key types of interactions as a starting point. With regards to A&IS these would include biological control, obligate fungal mutualisms, obligate host-parasite/parasitoid relationships, and obligate pollinator relationships. If the species interactions are known beforehand, it might be simply a matter of searching for both species on GBIF. However, ideally such data should be linked. Where species are tightly linked it would be important that such interactions are flagged, and when data are entered there is a way to indicate presence or absence of interacting species at a site.

Finally, while we feel that it is important that interaction information is included in GBIF, records of known interactions are currently housed elsewhere and this is probably the most appropriate approach. For example, the Global Biotic Interactions Database (http://www.globalbioticinteractions.org) provides records of which species are known to interact and on what basis (a species-level attribute), and Winston *et al.* (2014) compiles a list of organisms used in the classical biological control of alien plants. These databases are also potentially a source of new data on occurrences. The process would then be one of alerting people entering data, or exporting it that such interactions exist based on particular external data resources.

While we feel the lack of a tight link to data on species interactions is a major shortcoming of GBIF for research on A&IS, it is, of course, not an issue specific to biological invasions. We feel that if GBIF were to incorporate more data on interactions and facilitate its usage, there would be a qualitative increase in the value of GBIF as a general biodiversity information resource.

- Standards, such as Darwin Core, need to address the needs for documenting interactions.
 - GBIF contributors could be prompted to enter interaction data with recording systems and schemes for biodiversity observations encouraged to make more effort to record species interactions.
 - GBIF could facilitate access to fields containing interaction data and/or provide links between interacting organism records. Data on interactions could be flagged when data are extracted.
- GBIF could explore collaborations with current and new partners who house data on species interactions, with GIASIP playing a coordinating role with regard to interactions of specific importance to biological invasions
- GBIF could set up a test project to initiate work on species interactions. For example, GBIF could work with GIASIP and appropriate stakeholders (e.g. the International Organisation for Biological Control, <u>http://www.iobc-global.org</u>) to consolidate and mobilise data on the presence of biological control agents and host organisms.

9. Invasion impact information

Invasion impact information comes in a range of forms, including (i) freeform descriptions of impact and instances of impact, (ii) mechanisms of impact (e.g. predation or disease transmission), (iii) impact outcomes (e.g. negative impact on threatened taxa), (iv) allocated impact categories or classifications (e.g. minor, major, massive) (for ii-iv see Blackburn et al. 2015 and earlier related work, e.g. Nentwig et al. 2010, Kumschick et al. 2012), and (v) impact or prioritisation scores based on a risk assessment (e.g. Weed Risk Assessment). All of these forms of impact information are species-level attributes and are likely to be associated with an information source. Furthermore, we recognize that impacts vary geographically and through time, and the above examples are of aggregated information frameworks that are advancing for some taxa, especially in terrestrial systems.

Freeform impact information (i.e. (i)) is widely available via a number of existing platforms, such as the CABI Invasive Species Compendium and the Global Invasive Species Database, and the other forms of impact information are also likely to be available elsewhere (for example via EICAT, Blackburn et al. 2014).

There are many context, taxon and region-specific scoring and prioritizing systems for A&IS, and we do not recommend that GBIF interface directly with these. Rather, internationally supported and globally representative schemes or data platforms that provide impact information provide the most relevant basis for GBIF to include or to provide links to for users to access impact information on A&IS.

Accumulating the evidence of impact realised within a specific country or at a specific locality (i.e. as geographic or record-level impact attributes) will eventually provide high value information for A&IS research, policy and management (including at a global level, i.e. as species-level impact attributes).

Population level variables in individual observations provide the supporting information that is used to assess impact. Species interaction, accurate geo-locations, abundance and species status (as alien or native) all provide important information on the types of impact and their potential magnitude.

10. A&IS absence data

Absence data are particularly significant for A&IS, because they provide information essential to the prevention of A&IS introductions and risk management. Absence data may also be estimated as 'latest known date of absence'. Therefore encouraging the inclusion and estimation of absence records is an important part of providing data to support research on A&IS range dynamics (and to support A&IS policy and management). Population eradication records (equivalent to local extinctions of rare species) are also an important form of absence record for research on A&IS.

Efforts towards documenting, including and estimating presence and absence records with an associated date attribute will contribute significantly to improving the fitness for use of occurrence data on A&IS. We recommend that GBIF work with partners such as GEO BON to advance the needs in this area. Potential absences can uniquely be derived from "inventories", i.e. checklists, surveys or similar datasets with a specified taxonomic in addition to spatiotemporal scope (Guralnick et al., in review). Inventory-based non-detections are unique in allowing researchers to ascertain potential absences, either directly if they are "complete" or indirectly through modelling. This is in contrast with incidental records, where lacking information on taxonomic scope of sampling processes provides no way to infer absences without significant additional assumptions.

11. Prioritization of A&IS for policy and management

There is currently no single, widely adopted approach for prioritizing alien species for the purpose of ranking impact or decision support (the IUCN Environmental Impact Classification for Alien Taxa EICAT (Hawkins et al. 2015) is intended for this purpose). A recent overview concluded that the model or approach used varies across contexts and objectives (McGeoch et al. 2016). Nonetheless, as identified by Aichi Target 9, species-based prioritisation for biological invasions is an essential component of achieving global targets to reduce the size and impact of biological invasions. We recommend that GBIF contributed to the work towards the delivery of information integrated across species, pathway and site-based priorities. Importantly, improved information on species distributions and their native and introduced ranges is the most valuable contribution GBIF can make to improving the evidence base for species prioritisation and supporting pre- and post-border risk analyses for A&IS.

Information on abundance is important to assessing the magnitude of impact. Darwin Core captures abundance information in the three fields, individualCount and organismQuantity, combined with organismQuantityType. Although, these fields are quite simple they do provide sufficient information to inform impact assessments for many species. Nevertheless, surveys that qualify abundance, rather than occupancy, are rather rare.

Data and functionality to support prioritisation should be retained and improved. Prioritisation models use the type of data that are core to GBIF, including occurrence data (which combine information of species and site or area) and GBIF is therefore a potentially central resource for the process of prioritisation. Prioritisation models also use information on pathways and impact, and data structures that enable links between species, occurrence, pathway and impact information are especially valuable as input into prioritisation models for research, policy and management.

References

- Blackburn TM, Essl F, Evans T et al. (2014) A Unified Classification of Alien Species Based on the Magnitude of their Environmental Impacts. *PLoS Biology* 12(5): e1001850. <u>http://dx.doi.org/10.1371/journal.pbio.1001850</u>
- Carlton JT & Ruiz JA (2005) Vector science and integrated vector management in bioinvasion ecology: conceptual frameworks. In *Invasive alien species: a new synthesis*, Mooney HA et al., eds., 36-58. Washington: Island Press.

- Chown SL, Hodgins KA, Griffin PC, Oakeshott JG, Byrne M & Hoffmann AA (2015) Biological invasions, climate change and genomics. *Evolutionary Applications* 8(1): 23-46. http://dx.doi.org/10.1111/eva.12234
- Essl F, Bacher S, Blackburn TM et al. (2015) Crossing frontiers in tackling pathways of biological invasions. *Bioscience* 65(8): 769-782. http://dx.doi.org/10.1093/biosci/biv082
- Groom QJ, Desmet P, Vanderhoeven S & Adriaens T (2015) The importance of open data for invasive alien species research, policy and management. *Management of Biological Invasions* 6: 119-125. <u>http://dx.doi.org/10.3391/mbi.2015.6.2.02</u>
- Guralnick RP, Walls RL & Jetz W (2016) Humboldt Core toward a standardized capture of biological inventories for biodiversity monitoring, modeling and assessment. *Ecography*, in review.
- Hawkins CL, Bacher S, Essl F et al. (2015) Framework and guidelines for implementing the proposed IUCN environmental impact classification for alien taxa (EICAT). *Diversity and Distributions* 21(1): 1360-1363. <u>http://dx.doi.org/10.1111/ddi.12379</u>
- Kumschick S, Bacher S, Evans T et al. (2015) Comparing impacts of alien plants and animals in Europe using a standard scoring system. *Journal of Applied Ecology* 52(3): 552-561. http://dx.doi.org/10.1111/1365-2664.12427
- Latombe G, Pyšek P, Jeschke JM et al. (2016) A vision for global monitoring of biological invasions. *Biological Conservation*. <u>http://dx.doi.org/10.1016/j.biocon.2016.06.013</u>
- McGeoch MA, Genovesi P, Bellingham PJ et al. (2016) Prioritizing species, pathways, and sites to achieve conservation targets for biological invasion. *Biological Invasions* 18: 299-314. http://dx.doi.org/10.1007/s10530-015-1013-1
- McGeoch MA, Spear D, Kleynhans EJ et al. (2012) Uncertainty in invasive alien species listing. *Ecological Applications* 22(3): 959-971. <u>http://dx.doi.org/10.1890/11-1252.1</u>
- Nentwig W, Kuhnel E & Bacher S (2010) A generic impact-scoring system applied to alien mammals in Europe. *Conservation Biology* 24: 302-311. http://dx.doi.org/10.1111/j.1523-1739.2009.01289.x
- Pereira HM, Ferrier S, Walters M et al. (2013) Essential Biodiversity Variables. *Science* 339: 277-278. <u>http://dx.doi.org/10.1126/science.1229931</u>
- Revelle W (2014). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University.
- Ruiz GM, Fofonoff P, Hines AH, & Grosholz ED (1999) Nonindigenous species as stressors in estuarine and marine communities: Assessing impacts and interactions. *Limnology and Oceanography* 44(3:2): 950-972. http://dx.doi.org/10.4319/lo.1999.44.3_part_2.0950
- Ruiz GM, Fofonoff P, Steves B, Dahlstrom A (2011a) Marine crustacean invasions in North America: A synthesis of historical records and documented impacts. In *In the Wrong Place — Alien Crustaceans: Distribution, Biology and Impacts*, Galil BS, Clark PF & Carlton JT, eds., 215-250. Dordrecht: Springer. <u>http://dx.doi.org/10.1007/978-94-007-0591-3</u>

- Ruiz GM, Fofonoff PW, Steves B, Foss SF, Shiba SN (2011b) Marine invasion history and vector analysis of California: A hotspot for western North America. *Diversity and Distributions* 17(2): 362-373. <u>http://dx.doi.org/10.1111/j.1472-4642.2011.00742.x</u>
- Ruiz GM, Fofonoff PW, Steves BP, Carlton JT (2015) Invasion history and vector dynamics in coastal marine ecosystems: a North American perspective. *Aquatic Ecosystem Health and Management* 3: 299-311. http://dx.doi.org/10.1080/14634988.2015.1027534
- Smith B, Ceusters W, Klagges B et al. (2005) Relations in biomedical ontologies. *Genome Biology* 6: R46. <u>http://dx.doi.org/10.1186/gb-2005-6-5-r46</u>
- UNEP (2014) Pathways of introduction of invasive species, their prioritization and management. UNEP/CBD/SBSTTA/18/9/Add.1, subsidiary body on scientific, technical and technological advice, eighteenth meeting, Montreal. <u>www.cbd.int/doc/meetings/sbstta/sbstta-18/official/sbstta-18-09-add1-en.pdf</u>. Decision XII/17 CBD COP12
- Verloove F (2016) *Manual of the Alien Plants of Belgium*. Botanic Garden of Meise, Belgium. Accessed at: <u>http://www.alienplantsbelgium.be</u> on 25/08/2016.
- Winston, R.L., Schwarzländer M, Hinz HL, Day MD, Cock MJW & Julien MH, eds. (2014) Biological Control of Weeds: A World Catalogue of Agents and Their Target Weeds, 5th edition. Morgantown, West Virginia: USDA Forest Service, Forest Health Technology Enterprise Team, FHTET-2014-04.