GuardIAS

GUARDING EUROPEAN WATERS
FROM INVASIVE ALIEN SPECIES

Deliverable D1.1

G.I.A.S. - GuardIAS Intelligence Analyst System

AI-Based Algorithm to Query IAS Databases and Integrate Information

Planned delivery date: 30 April 2026 (M16)

Actual submission date: 4 May 2026

Deliverable leader: 17 - Nelson AI

Authors: Samuel Vander Velpen (Nelson AI) - lead developer, system architecture, programming and communication; Cris Novel (Nelson AI) - contributor

Internal Reviewers: Internal reviewers: Agnese Marchini (University of Pavia), Joost Andre Raeymaekers (NORD University), Jennie Pistevos (UAegean), Stelios Katsanevakis (UAegean)

Type of deliverable: New Concept and Tool (nCT)

Work Package: WP1 – Robust IAS data and innovative tools to underpin effective prevention

Work Package Leaders: Sergej Olenin (Klaipeda University), Anastasija Zaiko (Sequench)

Version: 1.0

Pages: 18

Funded by the European Union

GuardIAS receives funding from the European Union's Horizon Europe Research and Innovation Programme (ID No 101181413). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them.

GA No. 101181413

Start date of the project: January 1st, 2025

www.guardias.eu

DISCLAIMER

This document contains information on Guardias core activities. Any reference to content in this document should clearly indicate the authors, source, organisation, and publication date. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

COPYRIGHT NOTICE



This work by Parties of the Guardias Consortium is licensed under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). “

‘GuardIAS receives funding from the European Union’s Horizon Europe Research and Innovation Programme (ID No 101181413). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the EU nor REA can be held responsible for them.’

TABLE OF CONTENTS

EXECUTIVE SUMMARY **4**

ACRONYMS & ABBREVIATIONS **5**

1 INTRODUCTION **6**

2 METHODS **6**

2.1 DESIGN APPROACH 6

2.2 TECHNOLOGY STACK 7

2.3 DEVELOPMENT INFORMED BY GUARDIAS RESEARCHERS 7

2.4 THE NINE TOPIC AREAS 8

2.5 ENSURING RELIABILITY IN AI-ASSISTED EXTRACTION 8

3 RESULTS **10**

3.1 SYSTEM OVERVIEW 10

3.2 PIPELINE 1: DATA AGGREGATION 10

3.2.1 EU Distribution Map 11

3.3 PIPELINE 2: SOURCE FINDING (RESEARCH MODE) 12

3.4 PIPELINE 3: SOURCE EXTRACTION (RESEARCH MODE) 12

3.5 PIPELINE 4: REPORT GENERATION 12

3.6 COMPLIANCE WITH ORIGINAL D1.1 DELIVERABLE REQUIREMENTS 13

3.7 TESTING AND ITERATIVE DEVELOPMENT 14

4 CONCLUSIONS **15**

4.1 DIRECTIONS FOR FUTURE DEVELOPMENT 15

4.2 DIFFICULTIES AND RECOMMENDATIONS 16

5 BIBLIOGRAPHY **17**

6 ANNEXES **18**

TABLE OF FIGURES

TABLE 1. TABLE OF ABBREVIATIONS AND ACRONYMS USED IN THE DOCUMENT 5

TABLE 2. CORE TECHNOLOGIES USED IN G.I.A.S. AND RATIONALE FOR SELECTION. 7

TABLE 3. THE NINE TOPIC AREAS USED TO ORGANISE SPECIES DATA ACROSS ALL G.I.A.S. PIPELINES. 8

TABLE 4. OVERVIEW OF THE FOUR G.I.A.S. PIPELINES, THEIR MODE OF OPERATION, AND THEIR PRIMARY OUTPUT. 10

TABLE 5. DATABASES INTEGRATED INTO THE G.I.A.S. DATA AGGREGATION PIPELINE AND THE DATA EACH PROVIDES. 11

TABLE 6. MAPPING OF THE SEVEN ORIGINAL D1.1 DELIVERY STEPS TO THEIR IMPLEMENTATION IN G.I.A.S. 13

Executive Summary

This document reports on Deliverable D1.1 of the GuardIAS project: the development of G.I.A.S. (GuardIAS Intelligence Analyst System), an AI-powered multi-agent tool designed to significantly reduce the time researchers and practitioners spend gathering and processing invasive alien species (IAS) data.

IAS data is spread across numerous databases maintained by different institutions, stored in incompatible formats, and documented under inconsistent taxonomic names. Assembling a complete species profile manually is a multi-day process with a genuine risk of missing data due to the non-normalised way information is recorded across sources (Pagad et al., 2018). Beyond databases, relevant knowledge about IAS also exists in academic literature. Finding relevant papers and extracting specific facts relevant to a management question adds further time to an already demanding process. Both bottlenecks were consistently identified by GuardIAS researchers during development.

G.I.A.S. addresses both. It automates data aggregation from six trusted biodiversity databases, resolves taxonomic synonyms automatically, categorises all retrieved data into nine standardised topic areas, and generates formatted PDF reports with full source traceability. Directly informed by input from GuardIAS researchers during development, the tool was extended with a Research Mode that enables researchers to systematically search academic databases and AI-powered web search tools, extract verified facts from selected papers and merge those facts into the species knowledge base. Every step keeps the researcher in control of what is accepted and stored.

The tool has been tested with GuardIAS researchers and has developed in features and functionality based on that feedback. It is built on open-source components, adheres to FAIR principles, and will be hosted on LifeWatch and released as open-source software. Currently, the G.I.A.S. prototype is publicly accessible at: https://gias-prototype.streamlit.app/. Access requires a free Streamlit account; users can register at https://streamlit.io. The tool is currently open to all users. A thesis describing the tool's design and architecture in full will become publicly available, providing researchers and developers with a detailed technical reference.

Acronyms & Abbreviations

Table 1. Table of abbreviations and acronyms used in the document

Acronym / Abbreviation	Signification
AquaNIS	Aquatic Non-Indigenous Species Information System
D1.1	Deliverable 1.1 (WP1 AI Tool deliverable)
DOAJ	Directory of Open Access Journals
EASIN	European Alien Species Information Network
eDNA	Environmental DNA
FAIR	Findable, Accessible, Interoperable, Reusable
GBIF	Global Biodiversity Information Facility
G.I.A.S.	GuardIAS Intelligence Analyst System
GRIIS	Global Register of Introduced and Invasive Species
GuardIAS	Guarding European Waters from Invasive Alien Species
IAS	Invasive Alien Species
IUCN	International Union for Conservation of Nature
JSON	JavaScript Object Notation
LLM	Large Language Model
nCT	New Concept and Tool
TRL	Technology Readiness Level
WP1	Work Package 1
WRiMS	World Register of Introduced Marine Species

1 Introduction

Critical information about invasive alien species – distributions, ecological impacts, introduction pathways, management strategies – is spread across numerous databases maintained by different institutions. These databases store information in incompatible formats, use inconsistent field naming conventions, and index species under different taxonomic names depending on when each database last updated its taxonomy. The same organism can appear under multiple scientific names across sources, meaning a search by name alone will miss significant portions of available data (Pagad et al., 2018).

The practical consequence is that assembling a comprehensive species profile requires a researcher to manually search multiple databases, try multiple name variants for the same species, extract relevant information from each source individually, and reconcile wherever databases disagree. This process routinely takes days and carries a genuine risk of missing data due to the non-normalised way information has been documented across sources.

Beyond structured databases, a large body of relevant knowledge about IAS exists in academic literature. Finding the right papers, reading through them, and extracting the specific facts relevant to a management question is itself a time-consuming process. This was a need expressed directly by GuardIAS researchers during development: a systematic way to locate relevant academic papers and extract the information from them without having to do so entirely by hand.

G.I.A.S. (GuardIAS Intelligence Analyst System) was built to address both of these problems. It automates the aggregation of species data from multiple trusted databases and provides tools to find and extract information from academic papers, while keeping the researcher in control of every decision about what data to trust and include.

G.I.A.S. was developed primarily by Samuel Vander Velpen (Nelson AI), who led all system architecture, software development, prompt engineering, and design decisions throughout the project. Cris Novel (Nelson AI) contributed as a project participant.

2 Methods

2.1 Design Approach

G.I.A.S. was designed around three core principles, derived from the WP1 deliverable requirements and refined through consultation with GuardIAS researchers:

- Source traceability: every data point in the system carries a reference to the database or document it came from. No fact reaches the user without a traceable origin.
- Researcher control at every stage: no data is stored, merged, or included in a report without an explicit decision by the researcher to proceed. The tool handles time-consuming retrieval and pattern-matching; the researcher handles judgement.
- AI for interpretation, code for execution: deterministic tasks such as routing data, enforcing schemas, and deduplication are handled by purpose-written code. AI is used only where the input is unpredictable, such as correcting species names, mapping novel data fields, or locating relevant passages in free-text documents (Jordan & Mitchell, 2015).

The system is built as a multi-agent pipeline architecture. Multiple specialised AI agents, each optimised for a specific task, coordinate through a shared data cache rather than communicating directly. This keeps each agent's instructions focused, makes the system auditable, and allows intermediate outputs to be inspected and validated before the next stage runs (Guo et al., 2024).

2.2 Technology Stack

Table 2 summarises the core technologies used in G.I.A.S. and the rationale for each choice.

Table 2. Core technologies used in G.I.A.S. and rationale for selection.

Component	Technology and Rationale
User Interface	Streamlit (Streamlit Inc., 2024) and Plotly -- browser-based interface maintainable with basic Python proficiency; suited for open-source release to researchers
Pipeline Orchestration	Haystack AI 2.x -- native support for loops, parallel execution, and conditional routing (Deepset GmbH, 2024)
AI / LLM Provider	Mistral AI -- European provider operating under EU data protection law; appropriate for an EU-funded project (Mistral AI, 2025)
Data Storage	Session-isolated JSON file cache -- machine-readable, human-readable, matches API output format
Report Output	Pandoc + XeLaTeX -- formatted PDF output from structured data, no AI involvement in report assembly (MacFarlane, 2024)

2.3 Development Informed by GuardIAS Researchers

The original D1.1 specification defined the core function of the tool: aggregating IAS data from multiple databases into a single queryable system. During development, working sessions and feedback rounds with GuardIAS researchers revealed a consistent additional need. Researchers noted that academic papers contain important information relevant to IAS management, but finding relevant papers across multiple academic databases and extracting the relevant facts from them is a time-consuming process that is difficult to do thoroughly under normal working conditions.

This directly motivated the development of Research Mode, an extension to G.I.A.S. that goes beyond the original deliverable scope. Research Mode adds the ability to search for relevant academic papers across multiple databases and AI-powered web search tools, and to extract and verify specific facts from those papers in a controlled, traceable way. These additions were not

required by the original D1.1 specification. They were identified as essential through engagement with GuardIAS researchers and delivered within the same project phase.

2.4 The Nine Topic Areas

All species data across all four pipelines is organised into a fixed schema of nine topic areas, defined a priori in a central registry that acts as the single source of truth for category boundaries across all stages of the system. Table 3 lists these topic areas and their contents.

Table 3. The nine topic areas used to organise species data across all G.I.A.S. pipelines.

Topic Area	Content
Taxonomic Identity	Kingdom to species hierarchy, synonyms, taxonomic keys, author
Biological Traits	Morphology, physiology, behaviour, reproduction, lifespan
Distribution and Status	Native range, introduced range, establishment status per country
Habitat and Ecology	Habitat types, depth and temperature tolerances, salinity, diet
Species Interactions	Predators, prey, competitors, parasites, symbionts
Impacts	Ecological, economic, and human health impacts
Introduction Pathways	Vector and pathway categories, arrival dates; pathways are passed through as recorded in source databases, primarily following AquaNIS pathway terminology
Management and Biosecurity	Control methods, eradication attempts, biosecurity protocols, risk screening and risk assessment data, aquaculture relevance
Detection and Monitoring	Detection methods, eDNA protocols, early warning indicators

2.5 Ensuring Reliability in AI-Assisted Extraction

Because G.I.A.S. outputs can inform management decisions, preventing errors in AI-assisted extraction was a primary design concern. The system applies controls at multiple layers:

- When the AI identifies a relevant fact in a source document, it does not copy or paraphrase the text. Instead, a Python tool extracts the exact wording directly from the document by index position, confirmed through fuzzy matching. The stored value comes from the document itself, not from the AI's reproduction of it.
- After extraction, a Verification Agent runs three independent checks on each fact before the researcher sees the results. The first check asks whether the extracted value is actually present in the source text, catching cases where the AI has pointed to the wrong

passage. The second check asks whether the fact concerns the target species specifically, not a related species mentioned in the same paper; this matters because IAS literature frequently discusses multiple species together, and misattribution of a fact to the wrong organism would constitute a meaningful error in the data store. The third check asks whether the result is a reference to a figure or table rather than a direct finding; a pointer to data is not the same as the data itself and is not useful as a standalone extracted fact.

- After verification, the researcher sees every remaining fact alongside the source passage it was drawn from. They can deselect any fact they are not satisfied with before anything is merged into the data store. Nothing enters the knowledge base without an explicit researcher decision.
- Report generation draws directly from the structured JSON cache built during data aggregation and Research Mode. This approach greatly reduces the risk of AI-introduced errors in the final output, though no system can guarantee complete accuracy and researcher review remains essential throughout.

3 Results

3.1 System Overview

G.I.A.S. is organised into four sequential pipelines, as shown in Table 4. Each builds on what the previous produced, and the researcher sits at defined checkpoints between them to decide what moves forward.

The G.I.A.S. prototype is publicly accessible via a web interface at: https://gias-prototype.streamlit.app/. The application is hosted using Streamlit and requires users to log in with a Streamlit account. Registration is free and can be completed at https://streamlit.io. The tool is currently open to all users without restriction.

Table 4. Overview of the four G.I.A.S. pipelines, their mode of operation, and their primary output.

Pipeline	Mode	Primary Output
1. Data Aggregation	Default	Species data for nine topic areas aggregated across six databases; interactive dashboard
2. Source Finding	Research Mode (optional)	Relevance-scored list of academic papers from multiple databases
3. Source Extraction	Research Mode (optional)	Verified facts from approved papers, merged into the species data store
4. Report Generation	Default	Formatted PDF report with citations, EU distribution map, and regulatory status

3.2 Pipeline 1: Data Aggregation

The Data Aggregation pipeline is the core of the tool and the direct fulfilment of the D1.1 deliverable. The user enters a species name in any form – common name, misspelling, or outdated synonym – and the pipeline processes it in three phases.

First, the species name is validated and corrected using Mistral AI's language model, which draws on its training knowledge to correct misspellings and translate common names into accepted Latin binomials. This step does not query external databases; it uses the model's learned knowledge of taxonomic naming conventions. The corrected name is then used to query GBIF and WoRMS, which return lists of known synonyms. These synonyms are cleaned by a dedicated agent that filters out author annotations and retains only valid Latin binomials.

Second, the pipeline enters a synonym-aware loop: for each synonym, all six databases are queried simultaneously. This approach is essential because species names change over time and databases update their taxonomy at different rates, meaning a single-name search will miss

records catalogued under historical names (Pagad et al., 2018). Every raw API response is saved as a timestamped file for full traceability and is never modified.

Third, a hardcoded field mapping routes each returned data field to one of the nine topic areas defined in Section 2.4. This mapping was built by analysing real API outputs during development. For any field not covered by the mapping, an AI agent makes a one-off categorisation decision. The categorised data is organised into a session-isolated cache folder, with each data point tagged with its source database.

Table 5 lists the six databases integrated into the pipeline.

Table 5. Databases integrated into the G.I.A.S. Data Aggregation pipeline and the data each provides.

Database	Data Provided
GBIF (Global Biodiversity Information Facility)	Taxonomy, distributions, occurrence records, synonyms; also provides access to GRIIS country-level invasion data (GBIF Secretariat, 2026)
EASIN (European Alien Species Information Network)	EU regulatory listing, CBD introduction pathways, sector-based impact assessments (Katsanevakis et al., 2015)
AquaNIS	Aquatic neobiota, introduction records, invasion pathways (AquaNIS, 2026)
IUCN Red List	Conservation status, population trends, country-level presence, threat assessments (IUCN, 2026)
WRiMS / WoRMS	Marine species data, invasiveness classification, biological attributes, taxonomic backbone and synonym lists (WRiMS, 2026)
CABI Invasive Species Compendium	Directs users to expert-written species datasheets; subscription required for full content access (CABI, 2026)

All aggregated data is displayed in an interactive dashboard. Where databases return conflicting values for the same field, conflicts are highlighted prominently. The tool does not resolve conflicts automatically – that judgement remains with the researcher.

3.2.1 EU Distribution Map

As part of the species dashboard and the generated report, G.I.A.S. produces a Europe-focused choropleth map showing where the species has been recorded. The map is generated using the Plotly Python library directly from the distribution and status data collected during aggregation. Country-level occurrence records returned by the integrated APIs are parsed and classified – distinguishing native range from introduced or established range – and each country is colour-

coded accordingly. The map reflects only what the APIs have returned for that query; no additional geographic inference is made. If distribution data is absent for a species, the map is not included in the report.

3.3 Pipeline 2: Source Finding (Research Mode)

The Source Finding pipeline searches for academic papers relevant to the species and a selected research topic. It queries six sources in parallel: OpenAlex, Semantic Scholar, Europe PMC, DOAJ, Google Scholar, and Tavily. Tavily is an AI-powered web search tool that surfaces institutional and governmental sources alongside academic ones and can be toggled off when only peer-reviewed literature is required.

For standard research topics, curated search terms are stored in the system so the user does not need to know which keyword combinations work for each topic. For custom topics defined by the user, an AI agent produces a structured interpretation of what the topic means in the context of IAS research, covering what is in scope and what is not. The user reviews and corrects this interpretation before any searching begins, ensuring the system searches for what the researcher actually intends.

After results return, a Relevance Filter agent reviews the abstract of each paper and advises on whether it is likely relevant to the species and topic. This assessment is advisory only: low-relevance papers are flagged rather than removed, because the final decision on whether a source is worth pursuing remains with the researcher.

3.4 Pipeline 3: Source Extraction (Research Mode)

Once the researcher selects which papers to work with, the Source Extraction pipeline processes each one through a controlled three-phase workflow.

In the first phase, before any fact extraction occurs, the system converts the PDF to readable text and produces a brief overview of the study: where the research was conducted, which population was studied, and what methodology was used. It also scores which of the nine topic areas are likely covered by the paper. This overview is shown to the researcher so they can decide whether the source is actually useful for their purpose before committing to extraction.

In the second phase, the system extracts topic-relevant facts from the paper, with the AI identifying where each fact appears in the document. A Python tool then extracts the exact wording by index position. A Verification Agent subsequently runs three checks on each extracted fact: presence in the source text, correct species attribution, and whether the result is a direct finding rather than a reference to a figure. Facts that do not pass these checks are logged with the reason for removal and are not presented to the researcher as accepted results.

In the third phase, the researcher sees all facts that passed verification, each shown alongside the source passage from which it was drawn. They can deselect any fact before merging. Only facts the researcher actively accepts are written into the data store.

3.5 Pipeline 4: Report Generation

The Report Generation pipeline converts the accumulated species data from the JSON cache into a downloadable PDF. A cleaning step first collapses identical values that appear across

multiple sources and labels each unique value as consensus (supported by more than half of sources), single (one source only), or minority. This makes disagreements between databases visible without any AI involvement in resolving them.

The report is assembled directly from the structured cache data, with no AI generating report content from memory. This approach greatly reduces the risk of errors in the final output. The resulting PDF includes the GuardIAS logo, a species image, the EU distribution map described in Section 3.2.1, a risk summary, a distribution records table, regulatory status, and a complete references section listing database sources and academic papers separately.

3.6 Compliance with Original D1.1 Deliverable Requirements

The original D1.1 deliverable as specified in the Grant Agreement (also in Katsanevakis et al. 2024) committed to seven delivery steps. Table 6 maps each step to its implementation in G.I.A.S.

Table 6. Mapping of the seven original D1.1 delivery steps to their implementation in G.I.A.S.

Promised Delivery Step	Implementation in G.I.A.S.
1. Identify information sources and obtainable variables	Six biodiversity databases integrated via API (GBIF, EASIN, AquaNIS, IUCN Red List, WRiMS/WoRMS, CABI); nine topic areas defined a priori and consistently applied across all pipelines (see Table 3)
2. Standardise terminology and names across databases	Mistral AI corrects and standardises the input name before any API is called; a Synonym Cleaner agent filters valid Latin binomials from raw synonym strings returned by GBIF and WoRMS; a synonym-aware querying loop ensures all known name variants are queried across all databases
3. Homogenise data types and reconcile differences in scales and units	Hardcoded field mapping normalises predictable API fields; AI fallback categorises novel fields; a geographic normalisation service standardises place name references; a data cleaning step collapses duplicate values and labels consensus, single-source, and minority values across all databases
4. Optimise search queries by enhancing search efficiency, merging and synthesising results	Synonym-aware parallel querying maximises recall across all databases; per-database query formatting applied; curated search terms per research topic stored in a central registry; results deduplicated by DOI across all academic search sources
5. Integrate data into major IAS inventories (EASIN, AquaNIS, WRiMS, GRIIS)	EASIN, AquaNIS, and WRiMS integrated directly via API; GRIIS country-level data accessible through the GBIF integration; all four inventories represented in the data aggregation pipeline
6. Improve quality control with cross-validation algorithms	Three-layer extraction verification (source passage confirmation, species attribution check, figure-reference

	check); conflict highlighting in the dashboard; consensus/minority labelling in reports; a removed-fields log records every rejected fact with the reason for removal, providing a full audit trail
7. Deliver tool adhering to FAIR principles	Tool will be hosted on LifeWatch and released as open-source software; session-isolated JSON cache is machine-readable and human-readable; full source attribution on every data point; data accessible via API-first database integrations; a thesis describing the system design will be publicly available

3.7 Testing and Iterative Development

The tool has been tested with GuardIAS researchers, who used G.I.A.S. on real species of their own choosing and provided structured feedback. The tool has developed in features and functionality across multiple rounds based on this feedback, with improvements to species name resolution, academic search relevance, extraction verification, and session stability. The Research Mode itself was a direct product of this iterative process, having been identified as a priority need during early engagement with GuardIAS researchers.

4 Conclusions

Based on the development work carried out and the implementation described in this document, it is assessed that G.I.A.S. addresses all seven delivery steps committed to under Deliverable D1.1, and goes beyond the original scope in several areas. The tool reaches TRL 7 – an operational system tested with real users on real species – progressing from a TRL 2 starting point. The Research Mode, developed in response to needs identified by GuardIAS researchers rather than the original specification, represents a substantial addition to the tool's value within the same project phase. The system is publicly accessible via a web interface (https://gias-prototype.streamlit.app/), enabling immediate use and evaluation by external stakeholders.

4.1 Directions for Future Development

G.I.A.S. has established a working foundation with the potential to grow substantially in capability and impact with continued investment. The following directions are identified as most valuable, and could be explored with further cooperation and support:

- Country-specific and role-specific report outputs: different users in the IAS management chain have different information needs and different reporting obligations. There is potential to develop configurable output formats tailored to specific national regulatory requirements and roles, building on the data and categorisation infrastructure already in place.
- Risk screening and risk assessment integration: the nine topic areas already collect data relevant to risk screening and formal risk assessment processes. There is potential to structure report outputs specifically to support these processes, making the tool more directly useful at formal assessment stages of IAS management.
- More targeted and focused reporting: beyond generalised species profiles, there is potential to generate reports scoped to specific management questions, making the output more directly actionable for practitioners at different stages of the IAS response process.
- Expanded database coverage: additional databases can be integrated into the existing architecture. Priority candidates include those tentatively mentioned in the original proposal but not yet integrated.
- Horizon scanning and novel IAS identification: the accumulated species profiles could form the basis for a horizon scanning module, providing an early detection signal for species not yet present in Europe. This would connect directly to the objectives of WP2.
- Tighter downstream integration with WP2 and WP3: structured output formats and data access interfaces would allow G.I.A.S. outputs to feed more directly into species distribution modelling and early detection workflows, reducing the manual steps currently needed between data collection and downstream analysis.

The architecture of G.I.A.S. was designed with extensibility in mind, and none of the directions above would require rebuilding the system from the ground up.

4.2 Difficulties and Recommendations

The main technical challenges encountered relate to taxonomic instability – species accepted names change over time and databases update at different rates, meaning edge cases with recently reclassified species remain a known limitation and variability in API schemas across the integrated databases, which requires ongoing maintenance as those schemas evolve. PDF quality also varies for sources with complex layouts or non-standard formatting, which can affect extraction reliability. These are inherent challenges of working across heterogeneous external data sources rather than problems specific to the system design and are worth bearing in mind when setting expectations for data completeness.

5 Bibliography

AquaNIS. (2026). AquaNIS. https://aquanisresearch.com/

CABI. (2026). CABI invasive species. https://www.cabidigitallibrary.org/product/cb

Deepset GmbH. (2024). Haystack. https://haystack.deepset.ai/

GBIF Secretariat. (2026). GBIF technical documentation. https://techdocs.gbif.org/en/

Guo, T., et al. (2024). Large language model based multi-agents: A survey of progress and challenges. arXiv. https://arxiv.org/abs/2402.01680

IUCN. (2026). The IUCN red list of threatened species. https://www.iucnredlist.org/

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Katsanevakis, S., et al. (2015). European Alien Species Information Network (EASIN): Supporting European policies and scientific research. Management of Biological Invasions, 6(2), 147-157.

Katsanevakis, S., et al. (2024). GuardIAS: Guarding European waters from invasive alien species. Management of Biological Invasions, 15(4), 701-730.

MacFarlane, J. (2024). Pandoc: A universal document converter. https://pandoc.org/

Mistral AI. (2025). Mistral AI documentation. https://docs.mistral.ai/

Pagad, S., Genovesi, P., Carnevali, L., Scalera, R., & McGeoch, M. A. (2018). Introducing the Global Register of Introduced and Invasive Species. Scientific Data, 5, 170202.

Plotly Technologies Inc. (2015). Collaborative data science. https://plotly.com

Streamlit Inc. (2024). Streamlit. https://streamlit.io/

Vander Velpen, S. (2026). AI-powered multi-agent system for enhanced data collection and processing in invasive species management. Master's thesis, Victoria University of Wellington. [To be made publicly available.]

WRiMS. (2026). World Register of Introduced Marine Species. https://www.marinespecies.org/introduced/

6 ANNEXES

No annexes are included in this submission. The full technical documentation of G.I.A.S., including system architecture diagrams and pipeline descriptions, is provided in the Master's thesis by Vander Velpen (2026), which will be made publicly available upon completion.