CHAPTER 15

ARISE: a Dutch dataspace connecting nature and people

Elaine van Ommen Kloeke, W. Daniel Kissling, Julian Evans, Chantal Huijbers, Jacob Kamminga and Gerard Schouten

Abstract

Biodiversity is declining worldwide at an unprecedented rate. The densely populated country of the Netherlands is even one of the forerunners exhibiting this dramatic decline. In recent years, however, concern about the environment has moved decisively from niche to mainstream. In this chapter we introduce ARISE (Authoritative and Rapid Identification System for Essential biodiversity information), a government-funded research infrastructure that connects nature and people. The ambition of ARISE is to enable recognition of all natural species in order to monitor on a large scale end-to-end and near real-time biodiversity, thereby helping to 'bend the curve' of biodiversity decline. To do so, ARISE (i) provides an open data platform to collect field-captured samples and digital observations of all living organisms for species level recognition, and (ii) offers tools and services to challenge and engage researchers, policymakers and citizens to create the data to enable insights that help us to better understand biodiversity in relation to our environment and human activities. ARISE is linked to the Dutch national SURF infrastructure for data management and HPC capacity. Tools comprise: (i) an AI repository for species recognition models; (ii) smart annotation services for images and sound; (iii) dashboards, leaderboards and maps; and (iv) an array of sensors to capture multicellular species in their natural environment. As such, ARISE will be the 'one-stop-shop' or marketplace for species recognition services and non-invasive biodiversity monitoring.

15.1 Biodiversity at risk – what can we do to 'bend the curve'?

Biodiversity encompasses all kinds of life – the variety of animals, plants, fungi, and even microorganisms like bacteria that make up our natural world – as well as the diverse ecosystems they create through complex interactions between such living species. Humans are an integral part of this 'fabric of life'. Since the landmark paper of Hallmann *et al.* (2017) – which convincingly shows a 75% decline in insect biomass over the last three decades in nature protection areas in Germany – the topic of biodiversity has gradually shifted into the spotlights of policymakers and researchers. It has been estimated that over a million species are headed

 | DOI:10.3920/9789004730779_016

drivers of biodiversity loss

for extinction (Tollefson, 2019). As stated in the alarming Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (2019) report, the five major drivers of biodiversity loss are: (i) changes in land and sea use; (ii) climate change; (iii) pollution; (iv) direct exploitation of organisms; and (v) invasive species. These drivers are largely a result of anthropogenic activities (e.g. Kehou *et al.*, 2017).

Why is biodiversity important? To answer this question, we paraphrase from the booklet '*Biodiversity at risk: today's choices matter'* of the National Academies (2022): 'The strong and ancient connections between humans and other living species mean that we cannot really separate ourselves from the ecosystems within which we evolved. Although humans have come to dominate many of Earth's ecosystems, we still rely on these connections, making conserving and restoring biodiversity a matter of survival.' Biological processes make our planet liveable, mitigate climate change (Palmer, 2021, Hawkins *et al.* 2023), provide humans with food, clothing, medicines and building materials (e.g. Desborough & Keeling, 2017; IPBES, 2019); and finally, biodiversity protects our health and to a large extent supports our economies. One in five people rely on wild species for food and income. It has been estimated that the overall value of ecosystem services to human well-being equates to more than double the global gross domestic product (Costanza, 2014).

biological processes

values and behaviour

Our attitude towards nature is deeply rooted in our underlying values and behaviour. Western culture and philosophy have a long tradition of positioning humans above nature; people are encouraged and even obliged to control and exploit nature (Harrison, 2019). This attitude is counterproductive to halting biodiversity decline. Behavioural change is needed. In other words, a transition towards a nature-positive society that involves policymakers, conservationists, researchers, citizens, and organizations (Leclère *et al.*, 2020). The EU has acknowledged the urgency of this much needed transformation by articulating a biodiversity strategy[1] (as part of the Green Deal, a package of measures to make Europe climate neutral by 2050) aiming to (i) restore nature; (ii) improve the relationship between humans and nature; and (iii) reverse ecosystem degradation. A key success factor in implementing this strategy is reliable monitoring programs, i.e. biodiversity observation networks (Pereira *et al.* 2022; Gonzalez *et al.* 2023), and the development of so-called Essential Biodiversity Variables (EBVs). EVBs are essential for assessing, comparing and predicting the effects of human actions and spatial interventions (Navarro *et al.*, 2017).

ARISE

ARISE is the technical answer to large-scale monitoring; it provides an infrastructure and toolset to enable a cost-effective way of achieving the above vision. The ambition of ARISE is to become the central hub for all species recognition data and end-to-end services in the Netherlands. ARISE adopts an open science policy

1 https://environment.ec.europa.eu/strategy/biodiversity-strategy-2030_en.

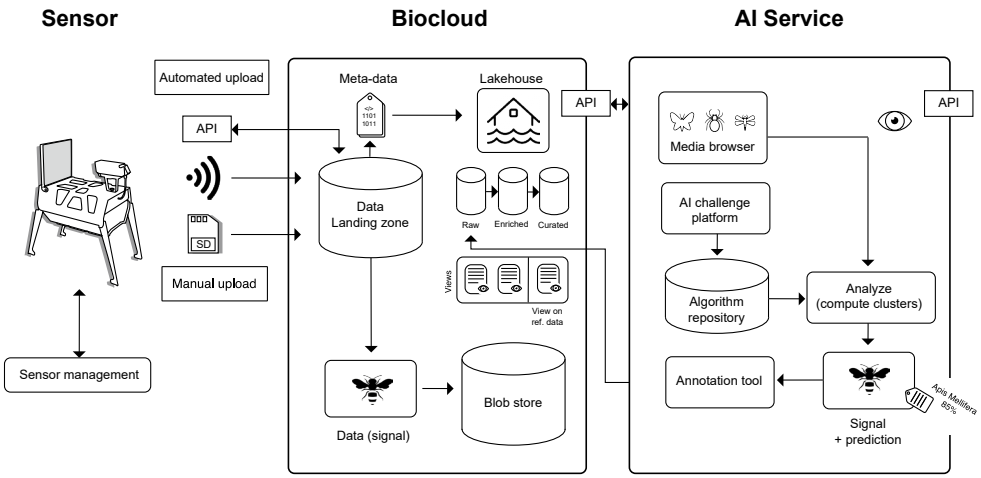

FIGURE 15.1 Overview of the end-to-end architecture in ARISE

(for data, code and publications), adheres to the FAIR data principles,[2] and respects data ownership by design. By building an *open* dataspace to store field observations (images, sound, radar) of all living organisms ARISE promotes and supports biodiversity research for everyone, including citizens. Biodiversity services derived from the data are developed and offered to researchers, policymakers, and companies, thereby connecting society with the challenge of biodiversity decline. Measuring the occurrence of species through time enables, for instance, estimation of soil and water quality through biological indicators. It also serves to answer ecological questions that are relevant for agricultural systems (such as the interplay between crop pollinators, pest species and their natural enemies). Furthermore, recognition of invasive species through sensor networks provides early warning signals for policymakers.

In this chapter we further detail what ARISE is and explain with tangible examples what automated species recognition for large-scale biodiversity monitoring entails. We start by sketching the landscape of measuring biodiversity in the field (Section 2). Next, as visualized in Figure 15.1, we focus on the end-to-end pipeline of (i) digital biodiversity sensors (Section 3); (ii) managing diverse and large amounts of data in a state-of-the-art data lakehouse system (Section 4); and (iii) digital species identification using AI, in particular deep learning (Section 5). We end this chapter with conclusions and lessons learned from the ARISE initiative so far and present an outlook for automated species recognition and biodiversity monitoring.

2 https://www.go-fair.org/fair-principles. FAIR stands for Findability, Accessibility, Interoperability, and Reuse of digital assets.

15.2 Measuring biodiversity in the field – a brief history and future outlook

observing species

It all starts in the field. Which creatures are out there, and how can we capture them with sensors? Understanding our natural living world and, in a way, measuring the state of biodiversity has been done for centuries. The first observations on species and their behaviour date back to the ancient Greek period, over time becoming more and more prominent in the age of enlightenment during the European Renaissance period through the naturalist movement ('History of biology', n.d.). Observing species in the field can tell us if a population is doing well or not, help us predict how an ecosystem is functioning or if we should expect trouble in the future for other species or even sense the effect of bigger underlying phenomena such as climate change. In this light let's have a look at one of the largest species groups that comprises biodiversity: insects. Insects comprise between 40–80% of biodiversity and perform all sorts of useful functions. For instance, they serve as food sources for many other species such as birds, fish and small mammals. They are also important for our food production system, as many crops depend on insect pollination, and they play an important role in nutrient recycling to keep the soil healthy. A drastic loss of insects can therefore have a serious effect on other species and even entire ecosystems. Recent studies have shown declines in insect populations in the Netherlands with 75% over the past three decades (Hallmann *et al.* 2017; Van Klink *et al.* 2020, Wagner *et al.* 2020). Keeping track of how well these species are doing is therefore more important than ever.

manual and labour-intensive methods

Collecting observations with the goal of accessing the state of biodiversity still relies on using manual and labour-intensive methods, ranging from on-the-spot observations by experts in the field, to setting traps, such as malaise and pan traps (Figure 15.2, top), in order to catch species in a particular area and time frame (Montgomery *et al.* 2021). These conventional invasive methods are used by long-running programmes, often through dedicated species foundations and ecological monitoring networks to answer trend-related scientific research questions. In the Netherlands one of the most well-known initiatives is NEM-net,[3] focussing on the International Union for Conservation of Nature (IUCN) red list of threatened species. This initiative has been monitoring since 1999, and data is stored in the national database for flora and fauna (NDFF) and by Statistics Netherlands (CBS in Dutch) for policy purposes. Additional data is also captured through more recently established species foundations or volunteering networks such as Observation International.[4] On an international level, observational data and other biodiversity-related data are captured through information portals, or so-called data repositories, with the Global Biodiversity Information Facility[5] (GBIF.org)

3 https://edepot.wur.nl/532548. NEM stands for the Dutch name 'Netwerk Ecologische Monitoring'. This means – roughly translated – Network for Ecological Monitoring'.

4 https://www.waarneming.nl or https://www.observation.org.

5 https://www.gbif.org.

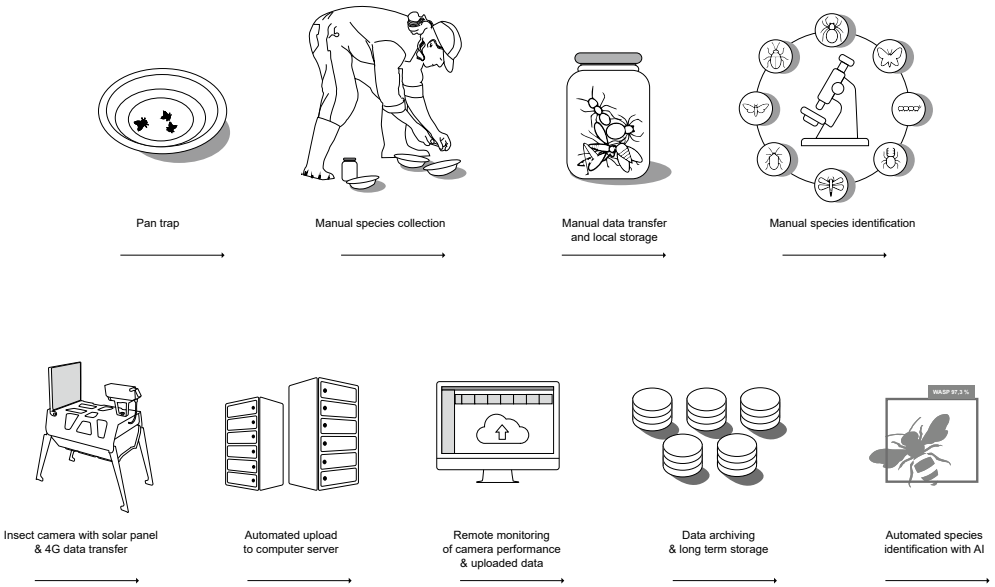

FIGURE 15.2 Manual data collection and identification (top) versus a fully automatic AI-powered solution for monitoring biodiversity (bottom)

being one of the largest with over 2.5 billion occurrence records, more than 90,000 datasets, 2130 publishing organizations, and an ever-growing number of peer-reviewed papers using the data (Saran *et al.*, 2022).

limitations

Even with all these long-standing and thorough initiatives it is currently impossible to know the full extent of our biodiversity. With the traditional methods we cannot answer the question 'what lives here?', as many species are inconspicuous, too small to be observed or in places not reachable for humans; so-called hidden biodiversity. One such place is the natural world underground, where almost all terrestrial plant life starts, including the crops we eat. It can also be the case that the human observer simply isn't in the right place at the right time or trapping methods attract some species more than others. Next to this, the traditional methods are usually extremely labour-intensive and hence can only be done in small areas, for a limited number of species and a few times per year (or less). This is where new technology can make a difference and complement the long-standing traditional methods through high-frequency observations, and at times that otherwise might not be feasible. Both DNA and digital sensor technology, capturing e.g. images, sound and radar, have the potential to capture species presence and answer all sorts of research and policy-related questions on a large scale and in semi-automated ways (Figure 15.2, bottom), and especially in combination with AI technology to make sense of the vast amount of data that these methods generate. ARISE can help us answer questions such as: Which prey are nesting birds eating this season? How is this species group doing compared to last year? Can we increase biodiversity if we change our agricultural practices or restore our ecosystems? Or even more

new technology

generally: What types of land use are adversarial or beneficial for biodiversity? Is an invasive species gaining in territory? More examples will be given in the following sections.

15.3 Digital biodiversity sensors – from smart nest boxes to insect soundscapes

Automated biodiversity monitoring with digital sensors is becoming increasingly feasible (Besson *et al.*, 2022). Compared to traditional biodiversity monitoring methods, stationary digital sensors such as digital cameras, microphones and radars allow high-frequency observations of species without observer disturbance, in remote areas or extreme environments, with relatively little labour and comparably low costs (Kissling *et al.*, 2018). Moreover, rapid advances in technology mean that sensors have become more affordable, can now be smarter and more autonomous, and remain in the field for longer periods of time. Within ARISE, the setting-up, running and maintenance of several digital sensors are being trialled across several monitoring demonstration sites within the Netherlands (current focus on three sites, see Figure 15.3). ARISE is thereby testing the use of digital sensors in a range of different habitats and conditions and demonstrating their innovation potential for biodiversity research and monitoring. The following sensors are currently deployed (as of October 2023):

15.3.1 *Sensors in the field*

Wildlife cameras: Two types of wildlife cameras are deployed. The Browning 2021 Spec Ops Elite HP4 triggered by passive infrared (PIR) represents a more traditional wildlife camera because it operates with batteries and requires data to be downloaded manually via an SD card. In contrast, the Snyper Commander 4G Wireless works more autonomously with a solar panel and 4G data transmission. It is either PIR triggered or operates in time-lapse mode and is deployed with either a wide (100 °) or regular (52 °) lens. The wildlife cameras are typically mounted near the ground (20–50 cm) to detect ground-dwelling birds and mammals (with PIR trigger), or on tall poles (e.g. 2 m) to count larger flocks or groups of animals (with time lapse).

Audio loggers: AudioMoth, a small, low-cost and full-spectrum audio logger is deployed for passive acoustic monitoring. It is capable of recording uncompressed audio of both audible and ultrasonic frequencies to microSD cards. The device is operated with batteries and can be configured with a certain sample rate and recording schedule. In ARISE, the audible frequency is currently used to record bird song and the ultrasonic frequency to monitor urban rats.

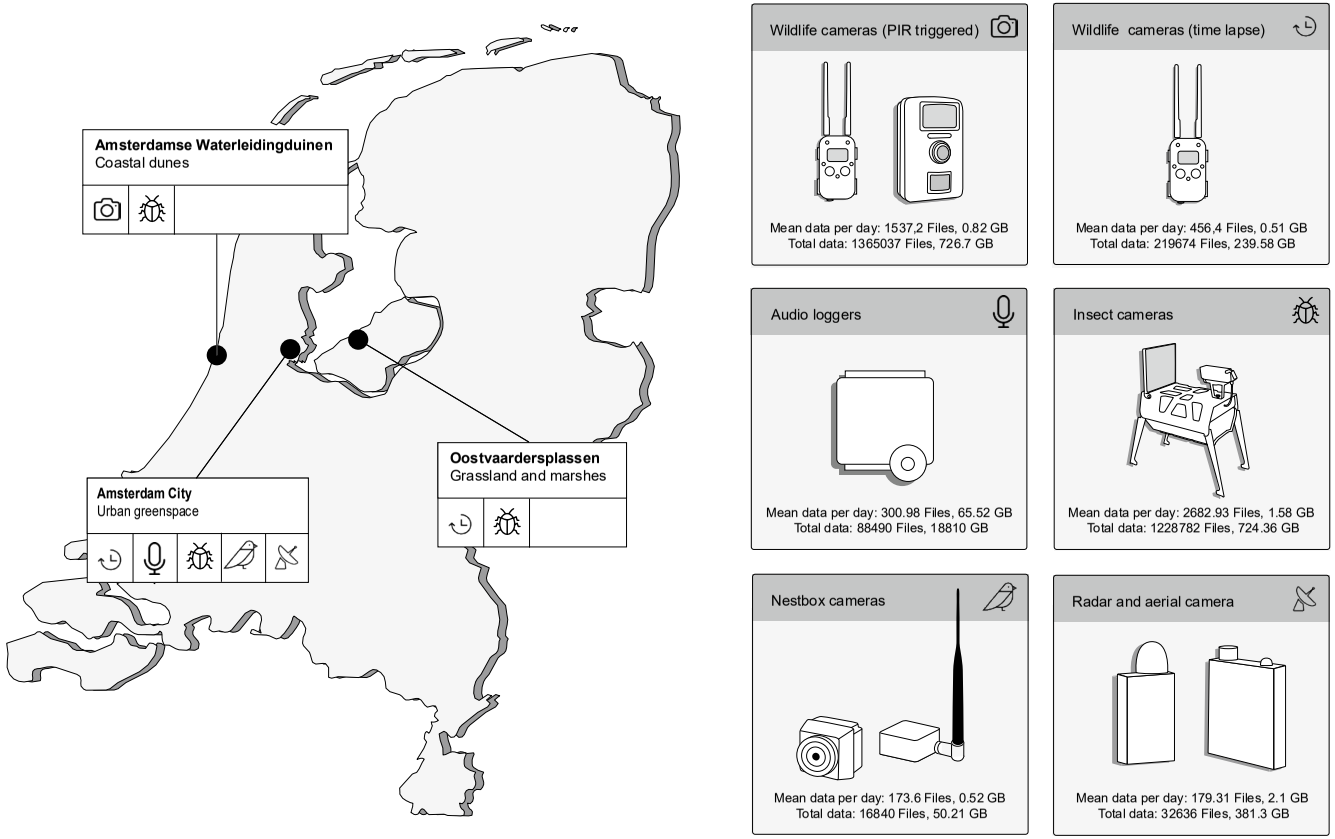

FIGURE 15.3 Diversity of digital biodiversity sensors tested in ARISE. (a) Location of the three ARISE monitoring demonstration sites in the Netherlands and the deployed sensors to monitor biodiversity non-invasively and remotely. (b) Different sensors and their data volumes

Insect cameras: Several insect cameras are deployed. This includes the solar-powered DIOPSIS camera[6] which attracts various groups of insects to a yellow screen, takes pictures when motion is detected and at regular intervals (every 10 seconds), day and night. Images and camera information can be sent to a server through the 4G network. Another insect camera in use is the Automated Moth Trap (AMT) which uses multiple light sources to attract moths at night (Bjerge *et al.*, 2021). The AMT version runs on mains power and within ARISE extra connectivity through a wired network was added to allow automatic data transmission. A third insect camera is a solar-powered insect detect camera trap[7] which uses an artificial flower platform for automated recording of flower-visiting insects (e.g. hoverflies). These insect cameras are assembled with low-cost off-the-shelf hardware components and can be combined with open-source software (Sittinger, 2022). For ARISE, the insect detection camera traps were modified to be solar powered and to transmit data automatically via 4G.

Nest box cameras: To monitor birds and the diet of chicks during the breeding season, intelligent nest box cameras were developed by a start-up company (Ecomoni[8]) in connection with ARISE. These cameras are motion-triggered through active infrared and automatically make a video recording when a bird enters the nest, for instance allowing the detection of food items that parents bring into the nest box for their offspring. The cameras can be operated with different lenses (63 °, 75 ° or 120 °) and currently work with batteries or mains power and use 4G or wifi for data transmission.

Radar and aerial camera: The Swiss Birdscan MR1 Radar[9] to monitor bird, bat and insect movements is also being tested through ARISE. The sensor uses a pulsed radar that emits beams vertically across a conically-shaped field to estimate the number of flying targets at an altitude of 30–1000 m (or up to 2000 m for large targets). Migration traffic rates for specific altitude layers are then computed. Wing beat characteristics of targets are used for target identification, which are typically groups of species that have similar wing beat frequencies (e.g. waders, songbirds, swifts, insects). The radar operates with mains power, and data transmission is through a wired network. To improve information on the species composition of the aerial fauna, a new and innovative aerial facing bird camera with a zoom lens (AeroecologyCam AC1) is used in conjunction with the Birdscan MR1 radar to gather image series of detected targets (e.g. small birds up to 400 m, large birds up to 1000 m). The AeroecologyCam AC1 also operates with mains power and data are transmitted through a wired network.

6 https://diopsis.eu/en/.
7 https://maxsitt.github.io/insect-detect-docs/.
8 https://www.ecomoni.nl/.
9 https://swiss-birdradar.com/systems/radar-birdscan-mr1/.

15.3.2 *Sensor autonomy, data pipelines and performance monitoring*

ARISE puts the emphasis on deploying sensors as autonomously as possible. Moreover, the development of automated data pipelines and the remote monitoring of sensor performance is a key focus. The currently deployed sensors broadly fall into three groups with increasing levels of autonomy: (i) traditional sensors, (ii) sensors with automatic transmission, and (iii) smart sensors. Traditional sensors such as the Browning wildlife camera and the AudioMoth require regular field visits to change the batteries and to manually collect data from the SD cards. To transmit the data from these sensors, which are currently some of the most commonly used sensors in ecological monitoring, ARISE has developed software to bulk download data from SD cards to a local storage device (e.g. laptop) and then upload the data to the sensor portal using an API. The sensors and their metadata must be registered manually. Moreover, sensor performance needs to be checked manually in the field to see if data are being correctly recorded.

Sensors with automatic transmission are more autonomous. For instance, the Snyper 4G wildlife camera is able to transmit data over 4G and can be powered with solar panels. This allows such sensors to remain in the field for extended periods of time with less human effort for repeated control and support. The data are transmitted over 4G to an external storage platform (e.g. FTP server or cloud service), from which they are automatically downloaded. Such cameras can also transmit a daily report of performance metrics, which allows users to remotely monitor battery status and on-board data storage over time. Problems can be remotely diagnosed, which reduces the amount of time in the field. These sensors and their metadata must still be registered manually in the sensor portal.

The most autonomous sensors are devices with on-board computational power capable of performing a variety of tasks. Most sensors deployed by ARISE fall into this category (e.g. insect cameras, nestbox cameras, radar and the aerial camera). These sensors can transmit data directly over 4G or wifi, rather than first transmitting them to an external repository. This allows them to attach a wide range of metadata directly to the media files, which can be automatically assigned to the correct sensor. The sensors' on-board computing typically results in high energy consumption which requires large solar panels (e.g. DIOPSIS system), large batteries (e.g. nest box cameras) or even mains power (e.g. Birdscan MR1 Radar, AeroecologyCam, AMT). However, they are more flexible in the kind of performance information they can transmit. This includes not only metrics such as battery status, but also information on CPU temperature or detailed logs produced by code running on-board. These smart sensors are capable of automatically registering themselves and their metadata to the sensor portal. They can also receive information from the infrastructure, allowing them to change settings remotely while the sensor is in the field.

15.4 All that data – managing a data lakehouse before it becomes a swamp

Innovative technologies such as digital sensors can greatly enhance our ability to conduct large-scale biodiversity monitoring and thus improve our understanding of which species live where. Yet, these technologies also produce vast amounts of data that need to be managed adequately for users to be able to access and analyse it. Deploying one sensor of each type described in the previous section would generate ~70 GB of data per day. As most studies would implement an array of sensors over an extended period of time, this quickly adds up to several terabytes per day (e.g. 10 sensors of each type = 4.2 TB/day). A comprehensive network of biodiversity sensors throughout the country could easily generate > 200 TB of data per month consisting of > 14M records.[10] Besides the actual media files, a data management system also needs to capture metadata such as information about the location, the sensors, deployments of a sensor, media type, projects, users, environmental information and other associated data (Figure 15.4). The combination of unstructured and structured data and the volume, variety and velocity with which data is generated means that the use of such technology, and therefore ARISE, needs a big data system to handle all this data.

data management technologies

The big data era has prompted a revolution in data management technologies (Harby and Zulkernine, 2022). Traditional data warehouses are sufficient for storing large amounts of structured data but are incapable of dealing with a large variety of unstructured data arriving at different velocities, leading to the implementation of data lakes over the last two decades (Khine and Wang, 2018). These data lakes, however, showed critical issues with processing data fast enough to prevent inconsistencies and errors resulting in unprocessed data in so-called data swamps. This initiated the development of a new solution in which desired features of data warehouses and data lakes are combined while addressing their weaknesses, resulting in the data lakehouse architecture which is rapidly becoming the industry standard (Armbrust *et al.*, 2021). This architecture provides low-cost storage in an open data format in combination with performance features such as indexing, caching, query optimization, and data versioning. The data lakehouse architecture has been implemented in other domains (e.g. biomedical research: Begoli *et al.*, 2022; maritime monitoring: Park *et al.*, 2023) and is also used by ARISE as a data management system.

10 Numbers based on 100 insect cameras, 100 wildlife cameras, 100 sound recorders and 5 bird radars deployed for 6 months.

15.4.1 *Biocloud: the brain that holds all ARISE information*

The Biocloud is the brain of the ARISE data management system (see Figure 15.1). It includes multiple layers, each serving a unique purpose in the data processing pipeline (Figure 15.4). The landing zone is our primary data intake layer. The Biocloud supports various data formats and utilizes a greedy ingestion process, meaning that all received data from a source is stored. The landing zone essentially acts as a staging area where the raw data is first collected before further processing. Once data is collected in the landing zone, the metadata is processed to the raw layer. Here, the data is stored in Delta tables in a Parquet file format, while its original structure is maintained. Delta tables provide the ability to perform update, delete, and merge operations on datasets, which can significantly improve the performance of the data processing pipeline (Armbrust *et al.*, 2020). After the raw layer, the data is moved to the enriched layer.

This layer acts as a transitional stage where the data is validated, cleaned and harmonized. The data is again stored in Delta tables, but it is now in a more refined, harmonized form that is ready for detailed analysis or further processing. This enrichment can involve cleaning up data anomalies, harmonizing disparate data formats, and enriching data with additional context or metadata. This results in one table per data entity (e.g. location, sensor, algorithm) that includes all data of that entity from different data sources. The last layer in our data pipeline is the curated layer. At this stage, the data is tailored into the format and shape that the end-users need. This can involve aggregating data, applying business rules, or transforming data to meet specific reporting or analytical requirements. The curated layer delivers data in a readily usable format, making it easier for business users, data analysts, and data scientists to extract insights without worrying about data clean-up or transformation tasks. Blob data such as images, sounds and videos, are also ingested

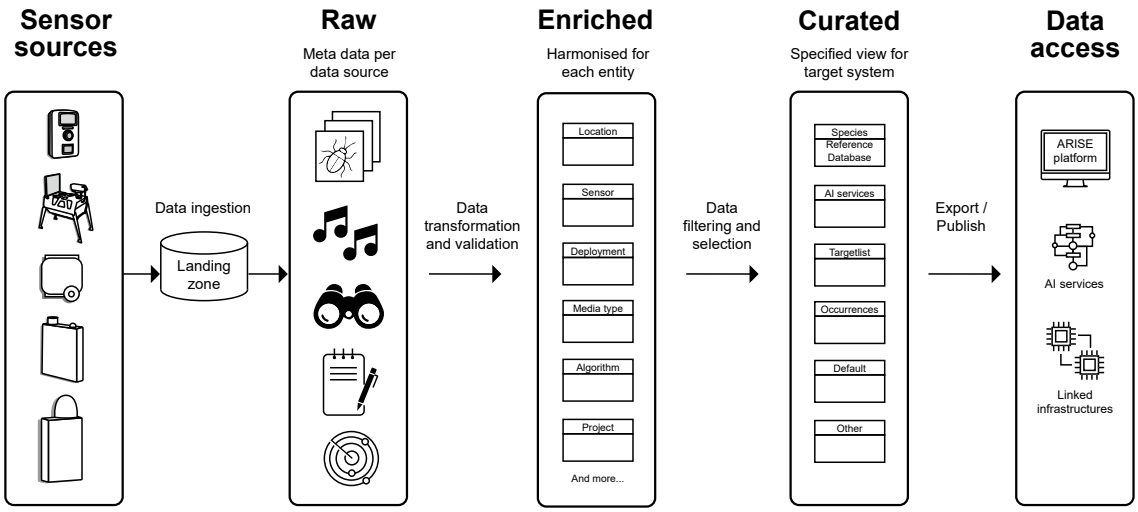

FIGURE 15.4 Overview of the Biocloud architecture with the different layers of processing the original data sources from raw to enriched and curated data for future use and access

into the landing zone, and then moved to S3 storage. The segregation between blob data and metadata allows for efficient querying and management of unstructured data while keeping the overall architecture optimized for performance.

Data stored in the Biocloud will be accessible through the ARISE platform as well as through machine-to-machine connections with other research infrastructures. The ARISE platform will provide a range of end user services such as search and browse through all publicly available ARISE data, view data from user-registered sensors, annotate data, and use AI algorithms to identify species from digital media (see also Section 5). While ARISE is mostly focused on services enabling species identification, other research infrastructures provide a wealth of information such as environmental data, human observations of species and global DNA barcodes, which can complement ARISE data for a broader perspective on the state of our ecosystems. Our goal is to connect with infrastructures such as the Global Biodiversity Information Facility (GBIF), Catalogue of Life, Barcode of Life Data System (BOLD), and Distributed Systems of Scientific Collections (DiSSCo) to contribute to the global assessment of biodiversity.

15.4.2 *FAIR by design*

The ARISE infrastructure is a complex system connecting and serving multiple systems, stakeholders, and user communities. Our design principles are aimed at building a system that is scalable and flexible, but also secure and reliable. Three of our guiding principles are:

*Adhere to the **FAIR principles:*** Data generated, analysed, stored and managed through ARISE can form the basis for many different avenues of biodiversity research. It is important to ensure provenance and full traceability of data objects so that analyses can be replicated, and data can be reused. Therefore, the architecture adheres to the FAIR principles as much as possible to make the data Findable, Accessible, Interoperable and Reusable (Wilkinson *et al.*, 2016). We implement Persistent Identifiers (PIDs), which are unique and unchangeable labels assigned to each digital object ensuring their accessibility and traceability over time (De Smedt *et al.*, 2020).

*Follow and use **international standards and practices:*** ARISE is dedicated to fostering open development by creating an extensible infrastructure that adheres to recognized standards. Data standards are key for consistency, efficiency and maintenance of high data quality by reducing the risk of errors and inaccuracies. Some well-known standards for biodiversity data include Darwin Core (Wieczorek *et al.*, 2012) and the Ecological Metadata Language (EML) (Fegraus *et al.*, 2005). More recently, GBIF presented a unified common data model supporting more complex types of biodiversity data,[11] and also more detailed standards for camera traps have

11 https://www.gbif.org/composition/7AZMWZtLvrnfYbFpUdF83I/diversifying-the-gbif-data-model-webinar-resources.

been published (Bubnicki *et al.*, 2024). ARISE will align with these standards as much as possible to promote interoperability and reusability of data.

Design for evolution: The ARISE infrastructure aims for a long life span; a time in which biodiversity research and technologies will evolve, so ARISE needs to be able to grow and evolve as well. We are building a system that will facilitate new types of sensors, data types, reference datasets or AI algorithms. The architecture is designed with scalability in mind allowing growth in volumes of data, number of data sources, users, service requests and connections with other infrastructures.

15.5 Powering AI for biodiversity

Digital biodiversity sensors generate vast amounts of data (see Figure 15.3), necessitating AI for analysis. ARISE enables non-AI experts to use AI for data interpretation and bridges the divide between computer science and ecology. Accurate algorithms are essential; errors in analysing biodiversity data can hinder crisis response. ARISE aims to identify every species in digital media, like images and audio, constantly improving accuracy. Furthermore, ARISE empowers researchers to evaluate, compare, and innovate AI for biodiversity. Its open, standardized approach ensures that breakthroughs are shared, adopted, and built upon, fostering a community of collaborative growth.

15.5.1 *The daunting challenge of AI in biodiversity*

Automatically identifying species using sensors (e.g. recording images or sound) and AI is a tremendous challenge. Biodiversity datasets have a long-tailed distribution and contain many examples of common species and few examples of rarer species. This constitutes a bias in the training of identification algorithms (van Horn & Perona, 2017) and is a crucial issue because rarer species are often the most important ones to monitor. Furthermore, there are millions of potential species to recognize, while the number of experts that can identify them is small and declining (Engel *et al.*, 2021; Greeff *et al.*, 2022).

The road to achieving the ARISE vision is not without its hurdles:

1. Quality data access: Ensuring high-quality training data to train high-performance algorithms
2. Dynamic digital framework: Crafting an adaptive digital species identification infrastructure that can evolve, allowing for updates and development of identification algorithms while preserving a clear provenance trail for automatic identifications
3. Employing modern AI: Developing the expertise and tools to leverage cutting-edge AI developments for species identification
4. Societal integration: Ensuring the infrastructure's relevance and upkeep in society for the long term

15.5.2 *The promise of ARISE*

The diversity of species contrasts sharply with the few experts capable of identifying them. Taxonomists and ecologists are therefore vital for integrating their knowledge into AI models. Ecologists frequently annotate excessive, non-essential data for AI learning, which could be more effectively focused on essential annotation tasks, like under-represented species. ARISE offers enhanced data annotation technologies and interactive AI training. ARISE aims to intertwine human expertise with evolving AI algorithms. While species observations were once manual, automation now aids detection and classification. Routine data can be accurately auto-classified, but complex data demands human validation to refine training data. In this collaborative model, as experts annotate, the AI simultaneously refines its capabilities as shown in Figure 15.5. With every iteration, AI handles broader datasets and highlights intricate data for human review. The objective is twofold: maximize expert efficiency and boost AI's proficiency in detailed species identification.

15.5.3 *Harnessing ARISE: the architecture behind advanced species identification*

three scenarios

Designed for researchers aiming to analyse biodiversity through digital sensors, the species identification architecture transforms heterogeneous 'raw' data into structured labelled data. When analysing data, a researcher typically faces one of three scenarios:

1. An available algorithm that performs adequately for fully automated reporting.
2. An available algorithm that requires enhancement or additional features (e.g. species granularity).
3. A lack of an algorithm tailored to the specific research goal.

While ideally only Scenario 1 would be needed for full automation, Scenarios 2 and 3 are commonly encountered, necessitating additional steps. As ARISE evolves, an increasing number of research objectives will fit into Scenario 1.

Using ARISE **involves:**

1. Connecting sensors or uploading media and organizing it.
2. Automatically analysing data for species detection and species identification, using:
    a. Existing AI algorithms, or
    b. New algorithms if current ones are unsuitable or outdated.
3. Finetuning, evaluating, and annotating new data or correct misclassifications.
4. Adjusting the algorithm for the specific domain and re-evaluating.
5. Iterating the above until satisfactory results are achieved.
6. Downloading final analysis results.

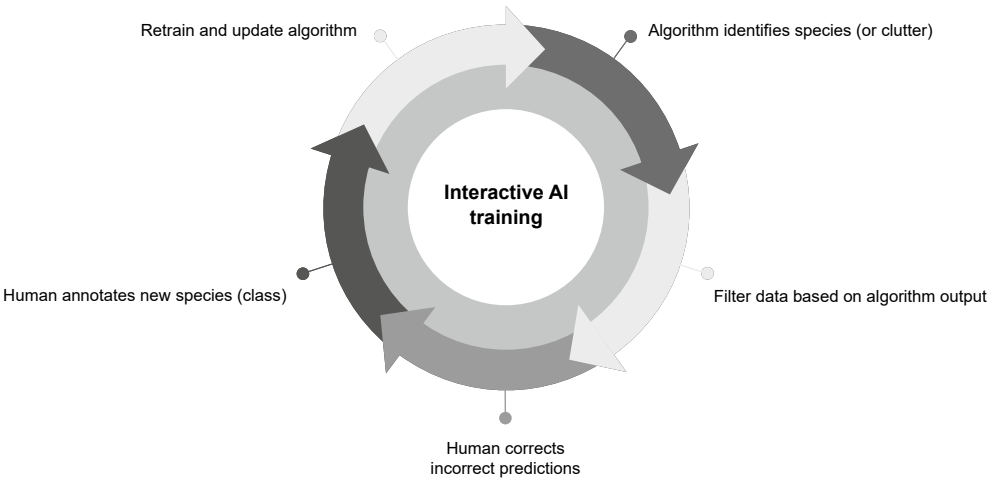

FIGURE 15.5 The active learning cycle of advanced species identification in ARISE

TABLE 15.1 Mapping of processes and scenarios to ARISE components

Process	Scenario 1	Scenario 2	Scenario 3	ARISE components
Data collection and storage	x	x	x	Sensor portal + Biocloud
Data visualization and organization	x	x	x	Media browser
Organization of algorithms and automatic deployment for data analysis	x	x	x	Algorithm repository
Data annotation for cleaning, training, and evaluation		x	x	Media browser + annotation tool
Algorithm evaluation and comparison		x	x	AI challenge platform
Algorithm design and training			x	AI challenge platform

process overview

Automated species identification processes, outlined in Table 15.1, remain consistent across media types, e.g. images and sounds. However, certain processes are specific to Scenarios 2 and 3 since they are automated in Scenario 1.

The key components of advanced species identification in ARISE are:

- *Media browser:* A central tool for experts to interact, visualize, label, and filter their data based on metadata like sensor ID, deployments, or analysis output. This component aids in creating datasets for analyses and annotations.
- *Annotation tool:* Beyond the media browser, ARISE integrates with external tools like Label Studio and Intel Geti, facilitating transitions to specialized annotation sessions and active learning. All annotations are synced with ARISE for data organization, training, and algorithm evaluation.
- *Algorithm repository:* ARISE hosts a range of open-source algorithms, from source code to automated model serving. Users can quickly deploy algorithms on platforms like Amazon cloud or SURF's Snellius. Transfer learning and fine-tuning principles enable existing algorithm reuse for related research objectives. If no suitable algorithm is found, ARISE's annotation tools assist in creating training data, and the AI challenge platform can be used to develop an algorithm.
- *AI challenge platform*: To accommodate evolving AI and diverse research needs, ARISE offers a challenge platform focused on biodiversity. Researchers can initiate and submit challenges, supplying necessary training and evaluation data acquired through ARISE. A dedicated evaluation container reviews submitted algorithms, with results displayed on a leaderboard. Successful algorithms can then be integrated into the algorithm repository for subsequent analyses.

15.6 Path forward and new challenges

As we embark on this journey towards autonomous biodiversity monitoring, the collaboration between technology, taxonomy, ecology, and biodiversity research has never been more crucial. In the Netherlands, ARISE offers a promising path forward, championing the cause of open science, expert involvement, multi-datatype monitoring, and the unparalleled capabilities of AI. The future beckons, and with tools like ARISE, we will deepen our understanding of the complexity of nature and have better tools to track progress towards achieving conservation management targets and biodiversity policy goals.

Although ARISE solves many of the issues related to large-scale non-invasive data collection as well as data interpretation by using AI, there are also new challenges ahead:

- How do we ensure that people start using ARISE? Building an end-to-end dataspace is not enough. How do we engage researchers, citizens and companies to use this vast amount of data to answer research, management and policy questions related to biodiversity?

– Is our architecture open enough? Can we easily add other types of sensors? For instance, drones that can automatically fly transects and capture photos at waypoints of a field with wildflowers. How many sensors do we need to estimate population sizes of species or the biodiversity that drives ecosystem functions and services?
– The advancement of novel technologies for understanding and measuring biodiversity should be driven by the scientific inquiries that researchers are tackling as well as by the requirements of society and policymakers. Do we have a process in place to capture (and implement in an agile way) these new requirements?

References

Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., … and Zaharia, M., 2020. Delta lake: high-performance ACID table storage over cloud object stores. In: Proceedings of the VLDB endowment, 13:3411–3424.

Armbrust, M., Ghodsi, A., Xin, R. and Zaharia, M., 2021. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In: Proceedings of CIDR, 8:28.

Begoli, E., Goethert, I. and Knight, K., 2021. A lakehouse architecture for the management and analysis of heterogeneous data for biomedical research and mega-biobanks. In: IEEE International Conference on Big Data, pp. 4643–4651.

Besson, M., Alison, J., Bjerge, K., Gorochowski, T.E., Høye, T.T., Jucker, T., … and Clements, C.F., 2022. Towards the fully automated monitoring of ecological communities. Ecology letters, 25:2753–2775. https://doi.org/10.1111/ele.14123.

Bjerge, K., Nielsen, J.B., Sepstrup, M.V., Helsing-Nielsen, F., and Høye, T.T., 2021. An automated light trap to monitor moths (*Lepidoptera*) using computer vision-based tracking and deep learning. Sensors, 21:343.

Bubnicki, J.W., Norton, B., Baskauf, S.J., Bruce, T., Cagnacci, F., Casaer, J., … and Desmet, P., 2024. Camtrap DP: an open standard for the FAIR exchange and archiving of camera trap data. Remote sensing in ecology and conservation, 10:283–295.

Costanza, R., De Groot, R., Sutton, P., Van der Ploeg, S., Anderson, S.J., Kubiszewski, I., … and Turner, R.K., 2014. Changes in the global value of ecosystem services. Global environmental change, 26, 152–158.

Desborough, M.J., and Keeling, D.M., 2017. The aspirin story – from willow to wonder drug. British journal of haematology, 177:674–683.

De Smedt, K., Koureas, D. and Wittenburg, P., 2020. FAIR digital objects for science: from data pieces to actionable knowledge units. Publications, 8:21.

Engel, M.S. Ceríaco, L.M., Daniel, G.M., Dellapé, P.M., Löbl, I., Marinov, M., … and Zacharie, C.K., 2021. The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. Zoological journal of the Linnean Society, 193:381–387.

Fegraus, E.H., Andelman, S., Jones, M.B. and Schildhauer, M., 2005. Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata

language (EML) and principles for metadata creation. Bulletin of the Ecological Society of America, 86:158–168.

Gonzalez, A., Vihervaara, P., Balvanera, P., Bates, A.E., Bayraktarov, E., Bellingham, P.J., … and Torrelio, C.Z., 2023. A global biodiversity observing system to unite monitoring and guide action. Nature ecology & evolution, 1–5.

Greeff, M., Caspers, M., Kalkman, V., Willemse, L., Sunderland, B.D., Bánki, O. and Hogeweg, L., 2022. Sharing taxonomic expertise between natural history collections using image recognition. Research ideas and outcomes, 8:e79187.

Hallmann, C.A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, … and De Kroon, H., 2017. More than 75 percent decline over 27 years in total flying insect biomass in protected areas. PLoS One 12;10:e0185809. https://doi.org/10.1371/journal.pone.0185809.

Harby, A.A. and Zulkernine, F., 2022. From data warehouse to lakehouse: a comparative review. In: IEEE International Conference on Big Data, pp. 389–395.

Harrison, P., 2019. Reformation of science. Aeon Magazine. https://aeon.co/essays/how-protestantism-influenced-the-making-of-modern-science.

Hawkins, H.J., Cargill, R.I., Van Nuland, M.E., Hagen, S.C., Field, K.J., Sheldrake, M., … and Kiers, E.T., 2023. Mycorrhizal mycelium as a global carbon pool. Current biology, 33:R560–R573.

History of biology. https://en.wikipedia.org/wiki/History_of_biology.

IPBES, 2019. Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Brondizio, E.S., Settele, J., Diaz, S. and Ngo, H.T. (editors). IPBES secretariat, Bonn, Germany.

Kissling, W.D., Ahumada, J.A., Bowser, A., Fernandez, M., Fernández, N., García, E.A., … and Hardisty, A.R., 2018. Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. Biological reviews, 93:600–625. https://doi.org/10.1111/brv.12359.

Kehoe, L., Romero-Muñoz, A., Polaina, E., Estes, L., Kreft, H. and Kuemmerle, T., 2017. Biodiversity at risk under future cropland expansion and intensification. Nature ecology & evolution, 1:1129–1135.

Khine, P.P. and Wang, Z.S., 2018. Data lake: a new ideology in big data era. In: ITM web of conferences ,17:03025. EDP Sciences.

Leclère, D., Obersteiner, M., Barrett, M., Butchart, S.H., Chaudhary, A., De Palma, A., … and Young, L., 2020. Bending the curve of terrestrial biodiversity needs an integrated strategy. Nature, 585:551–556.

Montgomery, G.A., Belitz, M.W., Guralnick, R.P. and Tingley, M.W., 2021. Standards and best practices for monitoring and benchmarking insects. Frontiers in ecology and evolution, 8:579193.

National Academies of Sciences, Engineering, and Medicine, 2022. Biodiversity at risk: today's choices matter. https://doi.org/10.17226/26384.

Navarro, L.M., Fernandez, N., Guerra, C., Guralnick, R., Kissling, W.D., Londono, M.C., … and Pereira, H.M., 2017. Monitoring biodiversity change through effective global coordination. Current opinion in environmental sustainability, 29:158–169.

Palmer, L. 2021. How trees and forests reduce risks from climate change. Nature climate change 11:374–377.

Park, S., Yang, C.S. and Kim, J., 2023. Design of vessel data lakehouse with big data and AI analysis technology for vessel monitoring system. Electronics, 12:1943.

Pereira, H.M., Junker, J., Fernández, N., Maes, J., Beja, P., Bonn, A., … and Zuleger, A.M., 2022 Europa Biodiversity Observation Network: integrating data streams to support policy. ARPHA preprints. https://doi.org/10.3897/arphapreprints.e81207.

Saran, S., Chaudhary, S.K., Singh, P., Tiwari, A., and Kumar, V., 2022. A comprehensive review on biodiversity information portals. Biodiversity and conservation, 31:1445–1468.

Sittinger, M., 2023. Insect Detect – software for automated insect monitoring with a DIY camera trap system. Zenodo. https://doi.org/10.5281/zenodo.7793296.

Tollefson, J., 2019. One million species face extinction. Nature, 569:171.

Van Horn, G. and Perona, P., 2017. The devil is in the tails: fine-grained classification in the wild. https://doi.org/10.48550/arXiv.1709.01450.

Van Klink, R., Bowler, D.E., Gongalsky, K.B., Swengel, A.B., Gentile, A. and Chase, J.M., 2020. Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances. Science, 368:417–420.

Wagner, D.L., 2020. Insect declines in the Anthropocene. Annual review of entomology, 65:457–480.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R. … and Vieglais, D., 2012. Darwin Core: an evolving community-developed biodiversity data standard. PloS One, 7:p.e29715.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A.,…. and Bouwman, J., 2016. The FAIR guiding principles for scientific data management and stewardship. Scientific data, 3:1–9.