SCOPING, HORIZON SCANNING AND PRE-SELECTING INNOVATIONS

Deliverable number: D2.1

Florian Rabitz[1]*, Rimantas Rauleckas[1], Rosalie van Dam[2], Alex Franklin[3], Sven Grüner[4], Veronika Kiss[5], Ilkhom Soliev[4], Elsa Tsioumani[6], Agnes Zolyomi[5]

1. Kaunas University of Technology
2. Wageningen University
3. Coventry University
4. Martin Luther University Halle-Wittenberg
5. GreenFormation
6. Transdisciplinary Institute for Environmental and Social Studies

* Corresponding author, email:

Florian Rabitz florian.rabitz@ktu.lt

June 2025

DAISY

DAISY receives funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101181857.

DAISY
LET'S TURN ON TRANSFORMATION

Project's technical information

Project acronym	DAISY
Project title	DigitAl, technologIcal and Social innovation mixes enabling transformation for biodiversity and equitY
Starting date	01st January 2025
Duration	36 months
Website	https://mydaisy.eu
Project coordination	Alex Franklin and Agnes Zolyomi

Deliverable's information

Deliverable number	D2.1
Work package no.	WP2
Deliverable title	Lists of innovations with key trends and issues
Task leader	Kaunas University of Technology (KTU)
Authors	Florian Rabitz (KTU), Rimantas Rauleckas (KTU), Rosalie van Dam (WR), Alex Franklin (CU), Sven Grüner (MLU), Veronika Kiss (GF), Ilkhom Soliev (MLU), Elsa Tsioumani (TIESS), Agnes Zolyomi (GF)
Dissemination level	Public
Type of deliverable	Report

Version management

Version	Status	Date	Authors/Reviewers
0.1	[Draft]	16/06/2025	Authors: Florian Rabitz (KTU); Rimantas Rauleckas (KTU)
0.2	[Draft]	19/06/2025	Reviewers: Sven Grüner (MLU); Alex Franklin (CU)
0.3	[Draft]	23/06/2025	Reviewers: Florian Rabitz (KTU), Rimantas Rauleckas (KTU), Rosalie van Dam (WR), Veronika Kiss (GF), Ilkhom Soliev (MLU), Elsa Tsioumani (TIESS), Agnes Zolyomi (GF)
1.0	[Final]	26/06/2025	Reviewers: Alex Franklin (CU), Sven Grüner (MLU), Lindy Binder (CU)

Recommended citation

Rabitz, F., Rauleckas, R., van Dam, R., Franklin, A., Grüner, S., Kiss, V., Soliev, I., Tsioumani, E., & Zolyomi, A. (2025). Lists of innovations with key trends and issues. (Report No D2.1). Project 101181857 – DAISY. Brussels: European Research Executive Agency.

List of abbreviations and acronyms used in this document

Acronym	Definition
AI	Artificial Intelligence
CU	Coventry University
eDNA	Environmental DNA
ESSRG	Environmental Social Science Research Group
GF	GreenFormation
KTU	Kaunas University of Technology
ML	Machine Learning
MLU	Martin Luther University Halle-Wittenberg
TIESS	Transdisciplinary Institute for Environmental and Social Studies
WHO	World Health Organization
WIPO	World Intellectual Property Organization
WP	Work Package

Background: About DAISY

DAISY - DigitAl, technologIcal and Social innovation mixes enabling transformation for biodiversity and equitY - will advance understanding of how specific mixes of interventions including social-technological innovations can be used to induce transformation for biodiversity and equity.

DAISY's main objectives

- To understand which socio-economic, political and behavioural processes, and their interrelationships, shape and enable our personal, political and practical ability to respond to the biodiversity crisis and how they impact on transformative change.
- To collect existing tools, processes, interventions and innovations that are conducive to triggering transformative change with the understanding of what enables them to address biodiversity loss and social inequity.
- To create intervention mixes based on existing tools and innovations and apply them in practice to induce transformation in all three spheres (personal, political, practical) to support biodiversity and equity prioritisation in decision- and policymaking.

Our case studies to test innovations

Innovation mixes will be tested and assessed for effectiveness in five seed innovation intensive case studies, within the domains of agri-food, education, energy and urban and regional development.

Turning on transformation

DAISY will have a special emphasis on amplifying innovation through bridging activities, networking events, wide stakeholder engagement and collection, connection and distribution of innovation seeds to switch on transformation.

Executive summary

This deliverable uses machine learning to generate a long list of social, digital, and other technological innovations with potentially transformative impacts on biodiversity and equity. We use topic modelling, an approach that identifies latent semantic structure across large bodies of text, to extract such innovations from the scientific literature as well as from patent applications, drawing on the Patentscope and SCOPUS databases. Using AI-based methods for labelling and classification, our long list of 987 items (169 drawn from patent data and 818 from scientific publications) provides a panoramic overview of relevant innovations that will serve as the starting point of the horizon scan that will be carried forward under the remainder of Work Package 2 (WP2).

Table of contents

Introduction

1. Purpose of the Deliverable

Task 2.1 develops a longlist of innovations that will, in subsequent tasks of Work Package (WP) 2, be reduced to a shortlist, subject to further in-depth analysis. To generate this longlist, the team from Kaunas University of Technology (KTU) is applying Artificial Intelligence (AI) and Machine Learning (ML) techniques to two bodies of literature, scientific publications, and patent applications. Scientific publications can contain indications for potentially transformative innovations that have been identified by researchers. Patent applications, with an intrinsic bias towards digital and other technological innovations, disclose information on inventions that are new, useful, and involve an inventive step, to grant exclusive (temporary) rights for commercial exploitation to the inventor. With both scientific publications and patent applications being produced in vast numbers, AI / ML techniques are suitable for dealing with large bodies of text while foregoing the need for qualitative, manual inspection (Gentzkow et al. 2019; Korinek 2023).

2. Context and Relevance

Task 2.1 has the objective of producing a longlist of social, digital, and other technological innovations with potentially transformative impacts for biodiversity and equity. This task is part of the WP 2 horizon scan, which is generally understood as a 'search process [...] at the margins of the known environment and possibly beyond it' (Amanatidou et al. 2012: 209). Horizon scanning has gained considerable popularity in recent years, as policymakers and other stakeholders struggle with fast-moving environmental, technological, and other global challenges. In the context of the World Health Organization (WHO), a recent horizon scan assessed scientific and technological changes with potential transformative impacts on global public health

(WHO 2022). A horizon scan of emerging issues affecting global trends in biological invasions identified, among others, globalisation tendencies in the Arctic region, and shifts in international trade policy, as major factors shaping introduction pathways (Ricciardi et al. 2017). A different horizon scanning exercise for global biodiversity conservation identified various emerging trends with important biodiversity implications, from genetically-modified mosquitoes to environmental releases of perfluorinated compounds to hydraulic fracturing (Sutherland et al. 2018).

More broadly, in the context of transformative change for biodiversity, horizon scans are crucial for detecting potential opportunities, in terms of social, digital, and other technological innovations early on, so that stakeholders can promote and leverage them for biodiversity-positive impacts. Scaling up innovations that emerge in specific niches may require active policy intervention (see Smith 2007). Thus, decision-makers and other stakeholders require an overview of the broader panorama of innovations that could, potentially, play transformative roles for biodiversity and equity. The need for foresight is furthermore accentuated by the rapid pace of technological change and the accompanying decision uncertainty. In such a context, horizon scans can provide broad-based awareness of the evolving innovation environment.

3. Scope and Objectives

The deliverable extracts biodiversity-relevant social, digital, and other technological innovations from two principal data sources: the scientific literature and patent applications. The purpose is to generate a comprehensive list (longlist) of innovations with potentially transformative implications for biodiversity and equity. For generating this longlist, we cast a wide net that will inevitably result in the inclusion of innovations that are **not** aligned with positive transformational change for

biodiversity and equity. However, the longlist constitutes the starting point of the horizon scan under WP 2. In the subsequent Task 2.2, this longlist will be narrowed down, and specific items filtered out to be evaluated for their transformative implications.

The data sources for this task cover two major origins of potentially transformative innovations. On the one hand, scientists produce insights that can facilitate transformational change for biodiversity and equity. The publication pressure in contemporary academic research all but guarantees that such insights will find their way into the published literature, from where it is, in principle, available for extraction. On the other hand, inventors create innovations that can, in principle, contribute to transformational change. Here, we capture those innovations for which inventors apply for patent protection, which may result in the granting of a temporal and geographically limited monopoly for inventions that are novel, non-obvious, and useful. Unlike for scientific research, the purpose of patent applications is typically of a commercial nature. By including patent applications within the scope of our search, we thus acknowledge that commercial research and development can and does provide outputs that may, in principle, facilitate transformative change for biodiversity and equity.

For compiling our datasets, we use two different databases. For scientific research, we draw on SCOPUS. SCOPUS is one among several academic databases, including legacy databases such as Web of Science; and new ones, such as Dimensions. While there is some variation in coverage between them, and while this variation also fluctuates over time, SCOPUS is a world-leading provider that covers a wide range of journals and publications, thus making it a highly suitable starting point for our analysis (see Gerasimov et al. 2024; Martin-Martin et al. 2021). For patent data, we draw on the Patentscope database of the World Intellectual Property Organization

(WIPO). While this is not the only patent database in the world, it is a pre-eminent platform with broad geographical and thematic coverage, hosted and curated by the United Nations agency, and tasked with the management and promotion of intellectual property rights. Together, SCOPUS and Patentscope offer a wide range of materials, from scientific and commercial research and development, to inform the generation of a comprehensive list of innovations to feed into the next steps of the horizon scan to be carried out under Work Package 2.

4. Structure of the Document

The text below sets out the methodology used for this deliverable, with a broader introduction, by discussing the role of topic models for large-scale text mining. Afterwards, we discuss the data sources that we use for extracting biodiversity-relevant innovations. We then continue to elaborate on the specific steps through which we generate innovations based on these data sources, and how we classify and label those innovations. We conclude this deliverable with a discussion of the various limitations and weaknesses of our approach.

5. Target Audience

This deliverable provides an unstructured longlist combined with a preliminary analysis of biodiversity-relevant social, digital, and other technological innovations. It may serve as a point of reference for future academic research on biodiversity-related innovations. As such, it may be of interest to academics engaged with foresight methods, including those who themselves conduct horizon scans related to biodiversity (e.g. Sutherland et al. 2018; Herbert-Read et al. 2022). The document may also be of relevance to policymakers and other stakeholders seeking a 'one-stop

shop' for inquiring into the broader technological panorama in the field of biodiversity. This deliverable is also critical for the DAISY consortium members who will continue with the horizon scan under subsequent tasks of WP 2. Finally, this deliverable provides a broad framework for leveraging text-mining techniques for horizon scanning, and thus may also be of methodological interest to academics and practitioners in the wider foresight field.

Methodology

1. Approach and Research Design

Our approach draws heavily on *topic models*. These are statistical models that have been widely used for the past two decades for identifying latent semantic structures across large bodies of text (e.g. Blei et al. 2003; Roberts et al. 2014; Eshima et al. 2024). Topic models identify clusters of expressions that have a high likelihood of being in a meaningful semantic relation. The starting point for topic models is thus co-occurrence: expressions that frequently appear together across different texts have a certain statistical likelihood of being related at a semantic level. Topic models have been used in a variety of contexts, including for the analysis of scientific journals (e.g. Blei and Lafferty 2007); open-ended survey responses (Roberts et al. 2014); or for the analysis of topic-network structures (Rabitz et al. 2021; Green 2022). What makes topic models useful is that they can reduce the dimensionality of large corpora of text; that is, they can reduce the semantic noise in a corpus to identify its structural features (Blei et al. 2003: 994).

Numerous versions of topic models have been proposed over the years. Correlated Topic Models can estimate the correlations between topics at the level of individual

texts, so that, for instance, we might know that a text that addresses topic A is *also likely* to address topic B, or *unlikely* to address topic C (Blei and Lafferty 2007). Structural topic models can incorporate document metadata, such as publication dates, to predict topic prevalence (Roberts et al. 2014). This approach can identify, for instance, if topics are more likely to occur in certain types of publications than in others, or whether their likelihood changes over time. Dynamic topic models explicitly account for time by allowing for changes in topic structure over time. Keyword-assisted topic models allow users to provide keywords prior to the analysis, which provides greater directionality for model estimation (Eshima et al. 2024).

Despite these differences, topic models are always based on two primary components. The first is a topic-term distribution that associates each expression that occurs in a corpus of text with each specific topic that is being estimated. A topic that semantically deals with, say, climate change would have strong associations with expressions such as 'coal' or 'carbon', but only weak associations with (for instance) 'spaceflight' or 'vaccines'. The second component is a document-topic distribution, which provides the extent to which a given document (e.g. a scientific article or patent application) is composed of a given topic. A report by the Intergovernmental Panel on Climate Change, for instance, would strongly reflect topics related to global warming, whereas a report on, say, hazardous waste would not.

Based on this approach, we here use topic models for two steps of our analysis. First, we estimate models that identify topics addressing *biodiversity-relevant innovations:* when factoring a corpus into multiple topics, many of those topics will reflect semantic content that is unrelated to the scope of the project. A corpus that consists of scientific literature on biodiversity may include, for instance, topics that address specific research methods. By zooming in on those topics that have a clear focus on

biodiversity-relevant innovations, we ensure a tight semantic fit with DAISY's overall objectives.

Second, using the document-topic distribution, we select the most highly ranked texts that are associated with each of those topics. In other words, after zooming in on the topics that have relevance for biodiversity and innovation, we then focus on the texts that are most strongly associated with these topics. Topic models allow for a comprehensive ranking of document-topic associations. The choice whether to select, say, the top five or the top ten documents in each topic is ultimately discretionary. Here, we choose such a quantity of top documents from each topic that we end up with a roughly balanced distribution (across topics, but also across our two corpora, respectively consisting of scientific articles and patent applications).

Using this approach on corpora that have been compiled with specific keywords targeting biodiversity and innovation allows us to extract relevant social, digital, and other technological innovations. At the same time, we wish to highlight that automated procedures such as topic models come with limitations, and model outputs need careful evaluation and curation by a human operator. While our approach allows for the extraction of innovation-related information from vast amounts of text, subsequent manual (qualitative) analysis is a key element of the overall analysis.

2. Data

As noted above, we use two datasets for this deliverable, one that incorporates scientific literature and one based on patent applications. For the first dataset, we query SCOPUS with a search string that ensures a thematic focus on biodiversity, and

a conceptual focus on innovation. For the former, we thus use (stemmed) synonyms for biodiversity:

[biodivers OR 'biological diversity' OR 'biologically diverse' OR 'ecological diversity' OR 'ecologically diverse' OR 'ecosystem diversity' OR 'genetic* diversity' OR 'species diversity' or 'habitat diversity']*

To capture the innovation dimension, we use:

[invent OR innovat* OR technolog*]*

As is common for text-mining approaches, there are trade-offs in play. Search strings that are overly specific risk excluding material that may be marginally relevant and may result in corpora that are too small to enable robust inference with machine learning methods. This is why we do not include explicit references to 'equity' (and synonyms). Search strings that are too general risk including wide swaths of thematically irrelevant material, thus introducing noise into the subsequent analysis. The search strings noted above appear to offer a useful compromise in this regard. Notably, we have chosen not to key our search strings explicitly to the four DAISY domains (agri-food, energy, urban and regional development, education). The reasons are twofold. On one hand, domain-specific keywords would further restrict the scope of the search and render corpora that are too small to allow for meaningful analysis. On the other hand, the relevance of potentially transformative innovations for biodiversity may not be obvious *a priori* but could possibly be established in a bottom-up manner during the analysis itself.

The search strings given above result in 31042 scientific publications. Setting a cut-off date of 2005 to cover the last 20 years, 28719 articles remain. This cut-off date

ensures that only (somewhat) recent publications feed into the horizon scan. However, the dataset is heavily skewed towards recent publications: from the entire 20-year dataset, approximately 44% of articles were published in the last 5 years, and 70% in the last 10 years. We construct our dataset based solely on the *abstracts* of these publications. We assume that thematically relevant information will be condensed in the article abstracts. Focusing on abstracts also reduces the noise in the model, as full-text articles contain large amounts of thematically irrelevant information, such as methodological or theoretical discussions. Full-text articles also require extensive pre-processing, for instance to remove literature lists, figure captions, and so forth. Abstract data thus offers a convenient way forward for our analysis (see Rabitz et al. 2021).

To build our dataset of patent applications, we use WIPO's *Patentscope* database. This is the most comprehensive global database, which includes information on patents filed under WIPO's Patent Cooperation Treaty, as well as those filed with cooperating national and regional patent offices, such as the European Patent Office. *Patentscope* thus offers a comprehensive picture of patent applications in high-innovation jurisdictions. These patent applications provide a window into innovation processes with potentially transformative implications for biodiversity. As noted above, the motivation behind the disclosure of patent information is commercial. One point to note is that patent applications as such do not necessarily give us information about feasibility: a patent application is an attempt by an inventor to gain a temporary monopoly for the commercial exploitation of an invention that may or may not turn out to be viable. Also, our dataset does not contain information on whether patent applications are ultimately granted, and whether the resulting patent grants are being upheld. In this regard, we would like to flag that the final model output does

contain patent applications that, from our point of view, amount to frivolous filings that are highly unlikely to result in patent grants.

Another important qualifier, and one that is impossible to work around with our methodology, is the disproportionate amount of Chinese patent filings. The rapid growth in Chinese patenting activity over the past decades can only partially be explained by research and development and is widely understood to encompass a broad range of low-quality patents that do not necessarily cover robust inventions (e.g. Chen and Zhang 2019). As patent quality is impossible to ascertain within the scope of this task, it is important to bear in mind the limitations of this approach. Below, we use a workaround to this problem by explicitly including additional non-Chinese patents in our analysis.

As patent applications deal with innovations as such, we use the first part of the search string given above (*[biodivers* OR 'biological diversity' OR 'biologically diverse' OR 'ecological diversity' OR 'ecologically diverse' OR 'ecosystem diversity' OR 'genetical diversity' OR 'species diversity' or 'habitat diversity']*), foregoing the second part to key the search for innovations. As before, we limit our search to patent applications filed in the last 20 years. We machine-translate non-English entries from the resulting list into English. The total number of entries in the corpus is 1751.

3. Analytical Methods

For each of our datasets, we use the resulting topic-term distributions to identify topics that could plausibly be associated with biodiversity-relevant innovations. We then extract the documents (article abstracts or patent applications) that are most strongly associated with these respective topics. As the scientific corpus is

approximately ten times larger than the patent corpus, and as we aim for an output of 1000 items on the longlist, in accordance with the DAISY project Grant Agreement, to ensure balance across the corpora we aim for approximately 900 items extracted from scientific articles and approximately 100 from patent data. These numbers, just as the cut-off dates of 2005, are ultimately subjective choices. There is no intrinsic reason for longlisting a total of 1000 items (rather than, say, 700 or 2000), just as there is no intrinsic reason for choosing the year 2005 over, say, 2010 or 2015. As with most social scientific research, these are choices that are ultimately discretionary, without hard criteria for selecting one number over another. We also consider the target number of approximately 1000 innovations to be commensurate with the wider project context, volume of relevant innovations, as well as the associated range of domains addressed by DAISY.

For patent applications, we run a model with 15 different topics. This choice of topic numbers offers a solid balance between exclusivity (the specificity of terms to topics) and semantic coherence (the degree to which top terms co-occur in documents). Some topics aggregate descriptors for the technical implementation of inventions. Other topics have a thematic focus unrelated to biodiversity-relevant innovations. However, 11 out of 15 topics are, in principle, relevant for the analysis and will be used to identify key literature that references biodiversity-relevant innovations. The excluded topics cluster expressions that are unrelated to biodiversity and equity, for instance referring to technical and engineering details without wider social and environmental context. The 11 topics selected are briefly summarised in Table 1 below.

Table 1: Selected topics from the patent corpus

Topic no.	Label	Keywords
1	Wetlands	plant, plants, planting, wetland, community, species, aquatic, landscape, submerged, diversity, area, planted, construction, lake, improved, different, biodiversity, comprises, structure, growth
4	Soil Health and Organic Nutrient Management	soil, fertilizer, organic, improved, root, process, biodiversity, carbon, water, waste, liquid, content, nutrient, tobacco, field, growth, ecological, compound, improving, solution
6	Wetland Water Purification Systems	water, ecological, system, body, floating, purification, wetland, treatment, aquatic, quality, tank, comprises, pond, bed, island, improved, ditch, environment, area, surface
7	DNA sequencing	biological, activity, sample, dna, sequence, step, marine, library, identifying, acid, gene, detection, microbial, primer, identification, biodiversity, sequencing, species, diversity, provides
8	Grassland Ecological Restoration	vegetation, soil, restoration, ecological, planting, region, recovery, grass, species, grassland, seeds, rate, steps, sand, seed, survival, following, area, comprises, technology
9	Microbial Pollution Treatment	treatment, pollution, biodiversity, environment, material, agent, control, effect, sludge, microbial, natural, following, bacteria, preparation, parts, composite, effectively, microorganisms, comprises, source
10	Urban Wetland Conservation and Restoration	ecological, area, biodiversity, habitat, system, protection, evaluation, land, comprehensive, environment, restoration, space, construction, function, wetland, urban, conservation, comprises, target, based

11	Artificial Fish Breeding and Aquaculture	fish, culture, breeding, artificial, fishes, medium, production, strain, discloses, application, provides, sea, technology, preparing, steps, resource, large, cultivation, biodiversity, simple
12	Agroforestry and Pest Management	planting, forest, tea, field, land, garden, control, rice, tree, biodiversity, natural, cultivation, pests, reduced, trees, improved, planted, farmland, management, insect
14	Forest Biodiversity Analysis	species, data, index, diversity, biodiversity, evaluation, distribution, model, information, obtaining, based, analysis, forest, value, ecosystem, target, determining, steps, comprises, calculating
15	Riverbank Ecological Restoration	river, channel, zone, ecological, water, bank, dam, riverway, area, restoration, system, arranged, aquatic, revetment, pool, biodiversity, habitat, deep, structure, belt

As noted above, we aim for approximately 100 items drawn from the patent corpus. With 11 topics selected for closer inspection, for each topic, we then extract the 11 documents that are most strongly associated in order to obtain an appropriate number of innovations (in line with the overall balance between the patent corpus and the scientific corpus, as outlined above). We generate short descriptors by using GPT o4.1-nano to summarise each patent application abstract in 50 words or fewer. For instance, patent application No. 114246119, 'afforestation method for recovering artificial forest', is summarised as: '*This afforestation method restores man-made forests by soil preparation, selective cutting, planting, and tending, maintaining ecological balance, preventing erosion, enhancing species diversity, and promoting soil health, ultimately reducing land rental costs and improving forest sustainability.*' Now consider patent application no. 116091923, 'forest landscape integrity

identification method based on moving window algorithm'. Our model summarises this patent as follows: '*A method using a moving window algorithm and penetration theory to recognize forest landscape integrity at pixel scale, incorporating density and connectivity indices, enabling quick, detailed mapping for sustainable management, biodiversity protection, and ecological decision-making.*'

We repeat the same analysis for our scientific dataset, this time running a model with 62 topics to account for the larger corpus size. As with the model for the patent corpus, this choice of topic numbers offers a balance between exclusivity and semantic coherence (see above). We focus on topics with associated top terms that indicate some degree of *applicability* to biodiversity and equity of underpinning social, digital, or other technological innovations. On that basis, we exclude (for example) all topics that exclusively reference biodiversity-related terms (e.g. 'species', 'families', 'taxa', 'genera') without concrete potential applicability. The idea behind this approach is that not every scientific study will contain, explicitly or implicitly, proposed innovations related to biodiversity and equity. In fact, we should assume that a majority of studies in our corpus consist of empirical analyses of biodiversity-related issues without specifically highlighting potential courses of action for biodiversity and equity. With this in mind, we also exclude topics that are applicable yet associated with negative biodiversity impacts.

The lists of social, digital, and other technological innovations with transformative potential for biodiversity and equity generated through topic modelling and further processing of patent applications and scientific publications, are available on Zenodo (see Annex). As noted, we aim to generate a list of approximately 900 innovations from the scientific publication corpus. As we have chosen 26 topics for closer inspection, this means that we base our analysis on the top 35 scientific articles

associated with each topic, extracting any innovations they reference in the text of the abstract. As above, by choosing the top 35 articles (across 26 topics), we aim at a proportionate distribution of innovations drawn from the patent corpus and the scientific corpus. In the case of the patents dataset, we used the titles of patent applications as descriptors for the respective innovations. As we do not have this option for scientific publications, we use an AI-based summary to develop short descriptors of semantic content.

To finalise the longlist, we go through four additional steps. First, to ensure greater balance in the coverage of patent data, we add 100 additional entries to the patent list drawn from non-Chinese patent applications. As noted above, this adds additional balance, considering widespread concerns over the high volume of low-quality Chinese patents, which do not necessarily track innovative activity, that have been generated over the past few decades. Second, we remove duplicates from the respective lists generated through each of our two topic models. These duplicates occur when single scientific articles or patent applications are strongly associated with more than one topic. This reduces the output to 169 entries for the patent corpus; and to 837 for the scientific corpus. For the latter, we also remove entries that do not amount to specific (potential) innovations but merely constitute scientific analyses without indicating concrete innovative solutions. In other words, we remove entries that do not contain any actionable content, and rather limit themselves to description or causal analysis. This brings the final list down to 818 items. The final longlist thus comprises 987 items, 169 drawn from patent data, and 818 from scientific publications. The full list can be found on Zenodo via the annex.

Third, we then generate some initial classifications of those entries. These classifications, which assign innovation categories (social, digital, and other

technological) as well as DAISY domains (agri-food, energy, education, rural and urban development), are to be understood as preliminary, and will be developed further under Task 2.2. We allow for multiple membership, so that single innovations can be of a social, digital and/or other technological type; and that they can belong to multiple domains simultaneously. For this automatic classification, which merely serves as an additional input for the work under Task 2.2, we use a series of highly specific keywords that correspond to each category. For the agri-food domain, for instance, we use 'food*', 'crop*', 'nutri*', or 'farm*'; for digital innovations we use 'digital*', 'artificial intelligence', or 'comput*'; whereas for social innovations we use keywords associated with behavioural changes such as 'planning', or 'govern*'. The full list of keywords can be found in the replication materials. We stress once more that this is a preliminary classification to be further developed under Task 2.2. Under that task, we will then only retain those innovations that we consider 'transformative'.

Fourth, in addition to these classifications, we also use an AI model to summarise the innovation-related content of the relevant scientific abstracts and patent applications. These summaries are included in the longlist, found in the annex. They serve only illustrational purposes and are intended to facilitate reader comprehension. Finally, for patent applications, we use the titles of the relevant inventions to refer to the respective innovations. For innovations extracted from our scientific corpus, we use an AI model to create titles. To illustrate: from the publication by Osathanunkul (2025), entitled 'eDNA tech tracks lethal jellyfish with CRISPR precision', we extract an innovation with the title 'CRISPR-Cas12a eDNA detection for jellyfish monitoring'. Our model summarises the abstract to that article as follows: 'The invention is a rapid, sensitive, and cost-effective CRISPR-Cas12a-based eDNA detection method for identifying the presence of Chiropsoides buitendijki jellyfish in coastal waters, enabling effective early warning and monitoring without the need for expensive

equipment.' Similarly, the article by Tejada-Gutiérrez (et al. 2025) on 'ForestForward: visualizing and accessing integrated world forest data from the last 50 years' is summarised as '[t]his invention is a comprehensive data integration and analysis system that curates over 4,400 forestry datasets into a large-scale data warehouse using ETL protocols and NoSQL technology, and a web platform 'ForestForward' that enables users to visualise and explore the temporal evolution of ecosystems', with the relevant innovation being described as 'ForestForward: Integrated Big Data Platform for Forest Ecosystem Monitoring'.

4. Limitations

Several general and specific limitations apply to the analysis carried out in this deliverable. As a general limitation, horizon scans and other foresight techniques are indispensable yet imperfect methods for grappling with potential, probable, or preferable futures. The identification of social, digital, and other technological innovations that *might* play transformative roles for biodiversity and equity is an endeavour fraught with uncertainty. This applies regardless of the specific methods chosen, whether these be primarily qualitative (such as scenario analysis or Delphi-based methods) or, as in the present deliverable, primarily quantitative.

The quantitative approach that we use here offers numerous advantages when analysing large corpora of text, yet suffers from intrinsic drawbacks. One drawback is the limitation of machine-learning algorithms when tracking semantic content. Our model identified scientific texts that refer to specific innovations, but also texts that simply assessed cause-and-effect relations without proposing any actionable interventions. The method we use turns up texts that discuss, say, eDNA-based methods for biodiversity monitoring, but also texts, for instance, dealing with the

empirical impacts of wind turbines on avian biodiversity. This is why, at key steps of the analysis, we had to manually adjust or filter model output. More broadly, the approach used here will inevitably capture (some) innovations without transformative potential for biodiversity and equity. Determining this potential will take place in later steps of WP 2.

Although not necessarily a limitation, an important qualifier relates to the multiple discretionary choices that underpin the analysis above. There is no intrinsic reason why we chose a timeframe of 20 years for our analysis, rather than, say, 15 or 25. The same applies to the indicative target of 1000 innovations, rather than 800 or 1500. It is important to bear in mind that the analysis above is not 'hard science', but rather casts a wide net to provide a sufficient informational basis for the subsequent steps under WP2.

Similarly, the classification scheme used in this analysis is, by necessity, superficial: assigning innovations to types (social, digital, other technological) and domains (agri-food, energy, education, rural and urban development) requires the manual inspection of semantic content by a human operator. The keyword-based approach used here can serve as a first approximation, but it is necessarily vulnerable to misclassification.

Conclusion

This deliverable contains a broad range of social, digital, and other technological innovations with potential transformative impacts on biodiversity and equity. It serves as a starting point for further analyses in subsequent tasks of WP 2. It also provides a broad, panoramic overview of the contemporary innovation landscape, and can thus

also serve as a database for researchers and stakeholders looking to survey the field of biodiversity-relevant limitations. However, the wider relevance of the analysis carried out under this deliverable only manifests in conjunction with further qualitative, in-depth assessments of transformative potentials. The qualitative analyses to be carried out in the remainder of WP2 will accordingly complement the automated, machine learning-based process implemented under Task 2.1. This will allow us to gradually narrow down the initial longlist to a more specific set of concrete social, digital, and other technological innovations that do have credible transformative potential for biodiversity and equity.

References

Amanatidou, E., Butter, M., Carabias, V., Könnölä, T., Leis, M., Saritas, O., ... & Van Rij, V. (2012). On concepts and methods in horizon scanning: Lessons from initiating policy dialogues on emerging issues. *Science and Public Policy*, *39*(2), 208-221.

Bhargavi, I., Pratap, A. R., & Sri, A. S. (2024, June). An Enhanced EfficientNet-Powered Wildlife Species Classification for Biodiversity Monitoring. In *2023 4th International Conference on Intelligent Technologies (CONIT)* (pp. 1-6). IEEE.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The annals of applied statistics*, 17-35.

Chen, Z., & Zhang, J. (2019). Types of patents and driving forces behind the patent growth in China. *Economic Modelling*, *80*, 294-302.

Eshima, S., Imai, K., & Sasaki, T. (2024). Keyword-assisted topic models. *American Journal of Political Science*, *68*(2), 730-750.

Gerasimov, I., Kc, B., Mehrabian, A., Acker, J., & McGuire, M. P. (2024). Comparison of datasets citation coverage in Google scholar, web of science, Scopus, Crossref, and DataCite. *Scientometrics*, *129*(7), 3681-3704.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535-574.

Green, J. F. (2022). Hierarchy in regime complexes: Understanding authority in Antarctic governance. *International Studies Quarterly*, *66*(1), sqab084.

Herbert-Read, J. E., Thornton, A., Amon, D. J., Birchenough, S. N., Côté, I. M., Dias, M. P., ... & Sutherland, W. J. (2022). A global horizon scan of issues impacting marine and coastal biodiversity conservation. *Nature Ecology & Evolution*, *6*(9), 1262-1270.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, *61*(4), 1281-1317.

List, J. A. (2024). Optimally generate policy-based evidence before scaling. *Nature*, *626*(7999), 491-499.

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*(1), 871-906.

Osathanunkul, M. (2025). eDNA tech tracks lethal jellyfish with CRISPR precision. *Ecological Informatics*, *86*, 103008.

Rabitz, F., Olteanu, A., Jurkevičienė, J., & Budžytė, A. (2021). A topic network analysis of the system turn in the environmental sciences. *Scientometrics*, *126*, 2107-2140.

Ricciardi, A., Blackburn, T. M., Carlton, J. T., Dick, J. T., Hulme, P. E., Iacarella, J. C., ... & Aldridge, D. C. (2017). Invasion science: a horizon scan of emerging challenges and opportunities. *Trends in Ecology & Evolution*, *32*(6), 464-474.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American journal of political science*, *58*(4), 1064-1082.

Schaffer, C., Elbakidze, M., & Björklund, J. (2024). Motivation and perception of farmers on the benefits and challenges of agroforestry in Sweden (Northern Europe). *Agroforestry Systems*, *98*(4), 939-958.

Smith, A. (2007). Translating sustainabilities between green niches and socio-technical regimes. *Technology analysis & strategic management*, *19*(4), 427-450.

Sutherland, W. J., Butchart, S. H., Connor, B., Culshaw, C., Dicks, L. V., Dinsdale, J., ... & Gleave, R. A. (2018). A 2018 horizon scan of emerging issues for global conservation and biological diversity. *Trends in Ecology & Evolution*, *33*(1), 47-58.

Tejada-Gutiérrez, E. L., Mateo Fornés, J., Solsona, F., & Alves, R. (2025). ForestForward: visualizing and accessing integrated world forest data from the last 50 years. *Database*, *2025*, baaf018.

West, M., & Baettig-Frey, P. (2025). Designing an effective and engaging augmented reality game for children to learn about biodiversity. *Educational Technology & Society*, *28*(1), 194-212.

WHO (2022). Emerging trends and technologies: a horizon scan for global public health. Geneva: World Health Organization.

Annex

The lists of social, digital, and other technological innovations with transformative potential for biodiversity and equity generated through the topic modelling and further processing of patent applications and scientific publications: https://zenodo.org/uploads/15756525.