NEGEM

Quantifying and Deploying Responsible Negative Emissions in Climate Resilient Pathways

Description of the NEGEM database

Horizon 2020, Grant Agreement no. 869192

Number of the Deliverable	Due date	Actual submission date
8.5	**31.01.2022**	**31.01.2022**

Work Package (WP): 8 – Vision and framework for pathways analysis

Task: 8.6. Creation of the open-access NEGEM database and feeding the NEGEM results to other relevant databases

Lead beneficiary for this deliverable: VTT
Editors/Authors: Lassi Similä

Dissemination level: Public

Call identifier: H2020-LC-CLA-02-2019 - Negative emissions and land-use based mitigation assessment

Document history

V	Date	Beneficiary	Author/Reviewer
1.0	2022-01-17	VTT	Lassi Similä
1.1	2022-01-25	VTT	Lassi Similä / David Reiner (UCAM), Gonzalo Guillén-Gosálbez (ETH)
1.2	2022-01-31	VTT	Lassi Similä
1.3, Revised version	2023-06-04	VTT	Lassi Similä

NEGEM

Partners

VTT – VTT Technical Research Centre of Finland Ltd, Finland
PIK - Potsdam Institute for Climate Impact Research, Germany
ICL - Imperial College of Science Technology and Medicine, United Kingdom
UCAM - University of Cambridge, United Kingdom
ETH - Eidgenössische Technische Hochschule Zürich, Switzerland
BELLONA - Bellona Europa, Belgium
ETA - ETA Energia, Trasporti, Agricoltura, Italy
NIVA - Norwegian Institute for Water Research, Norway
RUG - University of Groningen, Netherlands
INSA - Institut National des Sciences Appliquées de Toulouse, France
CMW - Carbon Market Watch, Belgium
UOXF - University of Oxford, United Kingdom
SE - Stockholm Exergi, Sweden
St1 - St1 Oy, Finland
DRAX - Drax Power Limited, United Kingdom
SAPPI - Sappi Netherlands Services, The Netherlands

Executive Summary

Background

NEGEM - *Quantifying and Deploying Responsible Negative Emissions in Climate Resilient Pathways* - is a Research and Innovation project funded by the EU Horizon 2020 (H2020) Programme. Its key target is to assess the realistic potential of Negative Emission Technologies and Practices (NETPs) and their contribution to climate neutrality, as a supplementary strategy to emissions mitigation. Work package (WP) eight aims at formulating vision and NEGEM framework for pathway analysis, including combining and compiling of results of other NEGEM WPs. Task 8.6 of the project, *Creation of the open-access NEGEM database and feeding the NEGEM results to other relevant databases* is responsible for managing and delivering all the quantitative and qualitative data produced by the NEGEM project and its WPs.

This deliverable presents a plan for building up the open access NEGEM database, and its key building blocks and design premises. It is aimed that the database is extensively exploited by external users to the project already during the project. After end of the project, the database presents the legacy of it.

Key elements of the database

NEGEM research is based on multiple disciplines and novel concepts, and data managed in the project consists of several data types and formats. Many textbook definitions for the term database share the concept of organized collection of data. In order to systematically build up such collection for a project like NEGEM, common ground for the applied solutions might be valuable, and it is presented in this deliverable. The proposed solution for NEGEM database will be implemented stepwise in collaboration with the project partners, building on the following key elements:

1. **User interface** to be implemented on the NEGEM website, allowing easy navigation, exploring and access for data put available. Primary use case will be based on scientific exploitation.
2. **Data content** of the database will be determined to meet the criteria on reproducibility of scientific publications and other data as specified in the Data Management Plan. Data Management Plan of NEGEM provides an overview on all the data produced in the project. According to it, the data generated in the NEGEM project can be divided in the following categories:
   - Data generated / used by mathematical models that the NEGEM consortium has assembled, e.g. the global MONET model for technology assessment and the global TIMES-VTT IAM with spatially explicit, quantitative NETP potential and environmental impact, and energy and greenhouse gas emissions impacts, and LCA (Life Cycle Assessments) models. The data used as input and output for this type of models is typically quantitative.
   - Data generated via methods typical in social sciences, such as interviews, workshops, focus group discussions, and surveys. The data in this category includes qualitative data recordings of the discussions (meeting the ethical standards on requesting permissions when needed), notes or transcripts. Also, quantitative data, such as analysis of survey answers with statistical methods, for example, can be generated under this area of research.

   Domain-specific dialogue between database developers and researchers in NEGEM on any relevant additional content relying on structure of the database will be continued to determine the exact data content of the database.

3. **NEGEM data repository** as a technical solution for permanent storing of NEGEM data. Importantly, this enables DOI numbers, search functions, permanent storing and similar functionalities important and required for scientific open access purposes. In this deliverable, a justified test case for application of *Zenodo* repository in NEGEM, is presented.

Based on the reasons of distributed management and execution of data collection efforts, the proposed approach for NEGEM database can be called a modular "collection" of databases, datasets and other sources. This is seen to support flexibility needed in multidisciplinary project divided in ten different WPs. For efficiency benefits, practical fluency, and partners' established conventions regarding data, the NEGEM database solution highly relies on application of available tools rather than building an all-encompassing solution from the scratch.

Exploitation and policy relevance

Building of the database is seen to especially support the scientific exploitation of the project's non-commercial outcomes. Thereby, the task contributes to science-based policy making in global, European, and national levels.

NEGEM database is open to any stakeholder or research groups or private person interested in the results. As the results also aim at high policy relevance, the exploitation groups connected to policymaking are even more directly targeted with other tailored outputs such as policy briefs, newsletters, dissemination events, and deliverables. To foster the exploitation of the data, these outputs are put easily accessible and findable via the same user web-based interface as applied for the database.

Open research and open data approaches can increase transparency of research as well as enable wider dissemination, re-use and exploitation of data. This can, in turn, promote the development of innovative new services, products and business models. At best, wider opening of data can ultimately contribute to better quality of research and thus, better decisions.

While open data in general can create unforeseen paths, synergies and innovations for utilization of data, especially, data in the following areas is seen to have potential in supporting policymaking related to NETPs:

- Data on technical and cost parameters, such as potentials and constraints as well as techno-economic, environmental, or other characteristics of NETPs, contributing to the relevancy and consistency of modeling efforts to determine sustainable pathways in meeting the Paris Agreement goals;
- Data on visions of role of NETPs in climate dimension in global, European, and Member State levels for creating roadmaps and scenarios supporting in developing climate change mitigation strategies;
- Data on perceptions of different stakeholder groups based on surveys, helping to determine efficient strategies to improving acceptance of sustainable NETPs.

Table of contents

1   Introduction

The call to make research data open is increasingly included in many global and European policy initiatives[1]. Developments of ICT technologies such as cloud services and data transfer capability have given rise to consideration of open data approaches in many research fields, including those relevant to negative emission technologies and practises (NETPs). Generally, open research and open data approaches can increase transparency of research as well as enable wider dissemination, re-use and exploitation of data. This can, in turn, promote the development of innovative new services, products and business models. At best, wider opening of data can ultimately contribute to better quality of research and thus, better decisions. As a specific measure to advance this target, NEGEM project has identified the call for creating a NEGEM database including relevant data. In the description of the action (DoA), containing the details of how the NEGEM project will be carried out, database is specified and referred as follows:

> This task [task 8.6] will be responsible for managing and delivering all the quantitative and qualitative data produced by the NEGEM project taken into account the GDPR requirement by the EU. In addition, NEGEM database will include the relevant public data on NETPs and its implementation taken into account the copy-right limitations. Task 8.6 will also feed NEGEM data to the other relevant databases, like the IIASA database, which is used in the IPCC work. Another example is the ECHOES database, which has created and collected EU and its MS level data on social science and humanities in clean energy transition. D8.5 will describe the data management, e.g. which repositories are selected, what kind of data is shown in those repositories and how to get an access or find the data.

Also, the database is referred to in task *Creation of the NEGEM framework and pathways*, which describes the framework that will guide how to formulate, analyse, monitor and disseminate visions, pathways, scenarios, and strategies aiming at reaching the Paris Agreement (PA) and Sustainable Development Goals (SDG goals) with or without NETPs. It is further stated that the framework includes a collection of WP1-WP7 data, which will be fed to an open-access database, which will be described in [Deliverable 8.5, this deliverable].

Progression of the project during its first 20 months has enabled to specify the general targets of the NEGEM database described in DoA and make them more concrete. In the specification work, lessons from an earlier ECHOES project are learned (see text box on p. 11). As NEGEM database will include all[2] the eligible data produced in the project, it is, by definition, inevitable that the content of the database develops all over the lifetime of the project. To some extent, minor additions are possible even after the official close date of the project. Exploitation of the data by any user of the open access database is welcomed by the consortium. Hence, a plan of the key building blocks, design premises and available tools for implementing the database throughout the project and after it, is needed, and it is systematically presented in this deliverable.

---

[1] https://digital-strategy.ec.europa.eu/en/policies/open-data
[2] Even if "all" data is mentioned in the project plan, in practice, some consideration is self-evidently needed, as there are a lot of intermediate working documents and versions of the data etc. with no added value for the database.

2　NEGEM database development

2.1　Definitions

In everyday language and scientific contexts, term "database" is commonly used in many meanings. Cambridge Dictionary defines the term as "*a large amount of information stored in a computer system in such a way that it can be easily looked at or changed*". Many textbook definitions share the concept of *organized collection of data*[1]. Such collection can include components such as schemas, tables, queries, reports, views, and other objects. Also, the variety of characteristics of databases is wide. There exist, for instance, different types of content (e.g. numbers, interviews, and bibliographical information), user interfaces, technological solutions and usage purposes for databases.

Databases in most "genuine" meaning can be seen widely including components referred in the above definition. Such a database would allow for making queries, and having relationships defined between tables in schemas, allowing searchability, organisation and reporting of the data.

In practice, more loose conventions in database terminology are applied in different domains of research than pure information technology[3]. For example, NEGEM deliverable 4.1 will detail a comprehensive, member state-specific database of negative emissions technologies[4], and NEGEM deliverable 4.2 targets at a database of biogeophysical parameters along with region-dependent economic parameters[5]. Another H2020 project, TradeRES, presents a database of scenarios including common data e.g., on policy scenarios, emission prices, commodity prices, projected cost and technical parameters of energy conversion and transmission technologies[6]. These examples demonstrate the practice of referring data collections organised e.g. in Excel spreadsheets as databases, even if not strictly meeting the definitions of database applied in computer science terminology or similar, more focused areas.

NEGEM research is based on multiple disciplines and novel concepts. Data managed in the project builds on diverse and consists of several data types and formats. The details of the result and input data in different parts of the project are being developed simultaneously by several research teams throughout the project. Considering this, NEGEM database builds on a primary target of making data collections available for scientific community, already during the 4-year-long project whenever feasible. For efficiency benefits, practical fluency, and partners' established conventions regarding data, the NEGEM database solution highly relies on application of available tools rather than building an all-encompassing solution from the scratch. Hence, NEGEM database can be called a modular "collection" of databases, datasets and other sources. Especially, individual parts of the NEGEM database are qualified with genuine database characteristics to varying extent. Nevertheless, many records to be included in NEGEM database, or the complete set-up as whole, can truly be qualified as "organised collection of data".

There is also a relevant term of "research data repository", *a database infrastructure that is set up to manage, share, access and archive researchers' datasets*[7] that is convenient to introduce for the NEGEM database development. In NEGEM context, especially, the archiving function becomes essential. Largely, existing and widely used repositories are applied for the repository implemented as a part of the NEGEM database.

---

[3] E.g. https://courses.aiu.edu/SYSTEMS%20DATABASES/SEC%201/SEC%201.pdf, Wikipedia, according to Similä & Koljonen 2017.

[4] https://www.negemproject.eu/wp-content/uploads/2021/06/Deliverable-4.1-NETP-database.xlsx

[5] https://www.negemproject.eu/wp-content/uploads/2021/10/Deliverable-4.2-Biogeophysics-database.xlsx

[6] see e.g. D2.1_TradeRES_DatabaseScenario_H2020-1.pdf

[7] What is Research Data Repository | IGI Global (igi-global.com)

2.2   Objectives

Development of the NEGEM database is guided by objectives defined by the NEGEM project. Furthermore, objectives reflected from general benefits of open data, and the initiatives of the European Commission (e.g. EC 2020, 2021) to response to them pave the way for the developments of the database.

First, according to project objectives specified in DoA:

- NEGEM database will manage and deliver all the quantitative and qualitative data produced by the project, taking into account the GDPR requirement by the EU.
- NEGEM database will include the relevant public data on NETPs and its implementation, taking the copyright limitations into account.

In addition to open access *publications*, allowing open access to *data* is among requirements for the NEGEM project in article 29.3 of the Grant Agreement. As there are well-established publication databases managed by scientific publishers with widespread conventions, the requirement on open access *data* is specifically tackled by NEGEM database. According to guidance from the EC, it is further specified that the digital research data generated and/or collected during the action as follows needs to be in compliance with Open Access for the following parts:

- underlying data: the data, including associated metadata, *needed to validate the results resented in scientific publications*
- any other data including the associated metadata, as specified in the Data Management Plan.

Developing NEGEM database allows an efficient tool to contribute to meeting the above objectives, and Data Management Plan developed and updated in the first reporting period (November 2020, November 2021) provides a solid outline for the work. To foster openness, the content of the NEGEM database is not limited to the scientific publications but including access to relevant project data even beyond this requirement.

Specifically, examples of data be stored to the NEGEM database include cost data and technical parameters on NETPs and related biogeophysical constraints such as land and water use. Examples of NETPs technologies covered include Bioenergy with Carbon Capture and Storage (BECCS), Direct Air Capture (DACC), afforestation, and Enhanced Weathering. For BECCS processes, for example, typical technical data are the energy efficiencies of the system, CO_2 capture rates, build rates, etc. Model-based result data based on NEGEM analyses include environmental impacts, and estimated costs and biogeochemical potential of NETPs. Furthermore, NETPs pathway-based result data include primary and final energy use by region up to 2100, GHG emissions and sinks by region, and associated costs for different NETPs scenarios. From public and stakeholder assessment point of view, data on surveys on public perceptions and acceptability of negative emission technologies provide central options.

Second, in response to the call for open data, the European Commission has adopted s.c. *FAIR principles* (Findable, Accessible, Interoperable, and Re-usable) originally published by Wilkinson et al. (2016) as guidance for the H2020 projects[8], including NEGEM. With detailed principles referred at the GOFAIR website[9], the H2020 manual defines the FAIR principles as follow, with digital assets being

---

[8] https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-guidelines-oa-covid-19_en.pdf
[9] https://www.go-fair.org/

- *Findable* when they are described by sufficiently rich metadata and registered or indexed in a searchable resource. Digital assets should be uniquely identified through the use of Persistent Identifiers that are globally resolvable (PIDs).
- *Accessible* when they can be obtained by humans and machines upon appropriate authorisation and through a well-defined and universally implementable protocol.
- *Interoperable* when they follow a formal, accessible, shared and broadly applicable format and when a language for knowledge representation is used.
- *Re-usable*, when rich metadata and documentation is provided that follow relevant community standards and provide information on provenance.

In the new Horizon Europe Programme (2021-27) of the EC succeeding H2020, the open data policies are further strengthened. In Horizon Europe, there is mandatory open access to publications and open science principles are applied throughout the programme[10].

To enable efficient exploitation for the NEGEM database, a tailored database solution is developed considering the needs of relevant user groups. Key components of the database solution include user interface, structures for data stored, as well as technical and software solutions necessary for database construction.

A large number of potential users and *exploiters* of the database are in principle eligible for the NEGEM database. DoA identifies users of the database as scientific community, policy makers, NGOs, NETP technology developers, "hard" and "soft" technology owners and providers, and uses of the database as value chain assessment, research, and business development. According to Open Access guidelines of H2020[11], as far as possible, projects must then take measures to enable third parties to access, mine, exploit, reproduce and disseminate (free of charge for any user) the research data. Considering this set-up, it is seen justified to build the NEGEM database as open for everyone. Hence, NEGEM database would not include a closed part for copyright-limited or otherwise restricted content. However, information about the research, codes, links, and similar tools available to validate the results can be included even in the public NEGEM database even in such cases. For private data sharing during the project, if needed, closed infrastructures in use for the consortium, are considered sufficient.

NEGEM database will also serve the target of exploitation of the project, as formulated in DoA. The nature of the NEGEM project is considered more scientific than commercial. Hence, building of the database is seen, especially, to support the scientific exploitation of the project's non-commercial outcomes. Therefore, scientific community forms a primary group to be considered in design choices of the NEGEM database. While it must be underlined that NEGEM database is open to everyone interested and the NEGEM results also aim at policy relevance, these exploitation groups are even more directly targeted with other tailored outputs such as deliverables, newsletters, and dissemination events. To foster the exploitation, also these outputs are put accessible and easily findable via the same user web-based interface as applied for the database.

---

[10] European Commission, Directorate-General for Research and Innovation, Horizon Europe, open science : early knowledge and data sharing, and open collaboration, Publications Office, 2021, https://data.europa.eu/doi/10.2777/79699

[11] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

There are several possibilities for technical implementations for generic data repositories serving the scientific purpose for the NEGEM database. Several initiatives by the European Union and scientific organisations for open protocols in this have been identified in the ECHOES H2020 project by VTT in consultation with its experts on knowledge services. Examples of potential data platforms designed for research data project include the EU-originated Zenodo, EUDAT, B2SHARE, B2SAFE, as well as Figshare (by UK based Digital Science & Research Ltd).

The background elaborated above, as well the work completed after the release of project plan, allows this document to take a step further in the development of the database. Consequently, key areas and questions discussed relevant for NEGEM database specification are divided in *user interface, data content and structures and technical implementation* as presented in Table 1. The approaches and key responsibilities are also included.

Table 1. Components of the database to be worked on.

Component	Specifications/ questions	Milestones	Main approaches
1. **User interface**	Design, visualisation, functionalities	Identification of needs, interface designed	Consultations between VTT and ETA to develop user interface
2. **Data content and structures**	Dataset and format specification on content to be included	Identifying relevant external data and internal data for the database	Consultations between VTT and key partners delivering content, utilising Data Management Plan
3. **Technical implementation of the database**	Interfaces, setting up data repositories etc.	Working solution up and running	Discussions and agreements between key partners, VTT & ETA on implementation work

BOX: ECHOES database and its implications for the NEGEM database

The project ECHOES, "Energy CHOices supporting the Energy Union and the Set-Plan", was completed in 2016-19. ECHOES is an EU H2020 funded project, and it involved 14 partners from eight European countries. The project results provide a deeper understanding of individual and collective energy-related choices and behaviour. Especially, the results unlock policy potential of an integrated social science and humanities (SSH) perspective that influence individual and collective energy choices and social acceptance of the energy transition in Europe.

NEGEM partner VTT acted as leader of a Work Package "Formulation of SSH database and SSH indicators" in the ECHOES project. The ECHOES database (Similä & Koljonen 2017, Correia et al. 2018) was built as an open access database, and it includes eligible ECHOES result data and some external data. ECHOES data collection methodologies involved: literature and policy reviews, meta-analyses, case studies, interviews and focus groups, psychological studies (experiments and correlational surveys), netnography study, and an international survey and choice experiment conducted across 31 European countries. ECHOES Database highlights the results from ECHOES projects in various ways. A data repository implemented on Zenodo repository is embedded in the database for permanent scientific exploitation of the data. The ECHOES database is available through link: https://db.echoes-project.eu/echoes/home.

In order to complete the ECHOES database, existing databases in the fields of SSH and energy were reviewed and different options explored. Especially, the following areas were studied:

- Requirements for open data from the H2020 point of view (e.g. permanence of storing, metadata) and restrictions for publishing data (e.g. privacy),
- Characteristics of datasets from different fields of science (e.g. variables, data types),
- Functionalities of the database (e.g. searches, visual tools) and design of user interface (buttons, menus, visual elements, etc. through which the database is used),
- Technical ways to implement the database and the chosen functionalities.

As lesson relevant for NEGEM database development, coordination of the data collection effort in highly multi-disciplinary project proved time-consuming and demanded intense dialogue between the partners in ECHOES. As another H2020 project, NEGEM project shares the requirements for open access data that were studied for ECHOES. Also, reviews made for data contents, functionalities, and technical alternatives in different existing databases can be used in NEGEM database development.

As for differences between the two projects, the ECHOES database was published for the public at the end of the project. This meant that exploitation of data for users external to the project was possible afterwards. NEGEM is a longer, 4-year-project, and it aims at prominent scientific exploitation for its data. To ensure data being up-to-date for the scientific purpose, the NEGEM database is built more stepwise and published partially already during the project. Also, the target group for NEGEM data is to larger extent in expert-oriented modelling. This implies that resource-efficient approach for the NEGEM database can rely more on existing data infrastructures for scientific purposes instead of starting from a scratch.

2.3   Schedule

In NEGEM database development, stepwise approach for implementation is chosen. This is justified by the fact that different parts of data eligible for open access are generated piecewise by different WPs during the project. Thereby, tailored approaches are applied in the work for increasing the capability of including intermediate data that becomes available during the project. According to procedure presented in Table 2, phases of the NEGEM database development include steps on mapping of data, basic architectures set up, and dialogue with the experts in different WPs. The dialogue between substance experts and database developers (mainly VTT and ETA) may include additional steps, such as workshop on this topic. The workshop could be followed by smaller, focused meetings with key persons of the Work Packages leaders and the database developers to specify the data content. With these steps, it is targeted to get the central components of the database implemented by the end of the second period of the project (M30). The process will be continued until the end of the project (May 2024), at the point of which the NEGEM database features a balanced legacy of the NEGEM project.

Table 2. Schedule of the NEGEM database development

Phase	Schedule	Responsible
Mapping data types and data content in NEGEM, and data platform options	M1-M20	VTT/WP10 in consultation with WPs, utilising Data Management Plan
Publication of Deliverable 8.5 with description of NEGEM database	M20 (January 2022)	VTT/WP8
Step-wise implementation of database components in dialogue with the substance experts to ensure appropriate solution for each eligible data	M20-M30	VTT/WP8, ETA, other partners
Step-wise addition of data when enabled by research	M20-M48	VTT/WP8, ETA (web interface development), other partners (data preparation)
Full-scale database launch, presenting the legacy of NEGEM	M48 (May 2024)	VTT

3   Data collections in NEGEM

Data Management Plan (DMP) of NEGEM is published several times during the project (first version in November 2020, with updates in November 2021, November 2022 and March 2024). DMP is a standard requirement for H2020 projects, and it *outlines how the research data collected or generated will be handled during a research project, and after it is completed, describing what data will be collected / generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved*. NEGEM approach for DMP builds on a detailed description of the issues by each Work Package, thereby, it is highly usable as tool for mapping types needed in database development and designing an open access database for NEGEM.

3.1   Classification of data
3.1.1   General classes

Among NETP literature and research and their progression in the NEGEM project, there are many potential pieces of which the NEGEM database can consist of. Furthermore, the variety of technical means on how different components can be put together for the NEGEM database, is extensive.

Based on a review conducted for earlier ECHOES H2020 project focusing on social science and humanities (SSH) and energy data oriented databases (Similä & Koljonen 2017), from user perspective, data eligible for storing and organising in databases takes several forms. These forms vary from mixed collections of files, link lists or websites, and documents, to the most developed "genuine databases" that allow queries, searches and organisation of the data for the end-users. According to the review, databases run by national statistical agencies or institutes established to conduct a certain persistently important statistical task in society, present examples on extensive data infrastructures with a focus on quantitative data. Scientific bibliographical databases, on the other hand, were typical examples of qualitative databases identified.

Database size can range from very general and wide databases maintained by statistical officials such as Eurostat or national agencies, to highly specific, small datasets developed for a single study. The size of a single database could be registered by indicators such as storage space required (Mb), or the number of sectors/variables included in the database. Size is relevant for the NEGEM database development through, for instance, the limits set by data depository providers.

Data type indicates what kind of data is stored in the database. As an example, as presented on the website of Oregon State University[12], the following classification (Table 3) is provided with regards to data types in general. The table is presented to provide an extensive reference and not all the categories of Table 3 apply to NEGEM data.

---

[12] http://guides.library.oregonstate.edu/research-data-services/data-management-types-formats

Table 3. Classification of data (source: Oregon State University)

Observational
Captured in situ
Cannot be recaptured, recreated or replaced
Examples: Sensor readings, sensory (human) observations, survey results

Experimental
Data collected under controlled conditions, in situ or laboratory-based
Should be reproducable, but can be expensive
Examples: gene sequences, chromatograms, spectroscopy, microscopy

Derived or compiled
Reproducible, but can be very expensive
Examples: text and data mining, derived variables, compiled database, 3D models

Simulation
Results from using a model to study the behavior and performance of an actual or theoretical system
Models and metadata, where the input can be more important than output data
Examples: climate models, economic models, biogeochemical models

Reference or canonical
Static or organic collection [peer-reviewed] datasets, most probably published and/or curated
Examples: gene sequence databanks, chemical structures, census data, spatial data portals

3.1.2   NEGEM data classification

To build a sound solution for the NEGEM database, an initial categorisation for the NEGEM data to be generated in the project, is needed. That is, there are not off-the-shelf solutions available for such a wide and multidisciplinary project as NEGEM. On the other hand, the Work Package specific description does not meet the criteria of being convenient for external users. According to DMP, the data generated in the NEGEM project can be divided in the following categories:

- Data generated / used by mathematical models that the NEGEM consortium has assembled, e.g. the global MONET model for technology assessment and the global TIMES-VTT IAM with spatially explicit, quantitative NETP potential and environmental impact, and energy and greenhouse gas emissions impacts, and LCA (Life Cycle Assessments) models. The data used as input and output for this type of data is typically quantitative.
- Data generated via methods typical in social sciences, such as interviews, workshops, focus group discussions, and surveys. The data in this category includes qualitative data recordings of the discussions (meeting the ethical standards on requesting permissions when needed), notes or transcripts. Also, quantitative data, such as analysis of survey answers with statistical methods, for example, can be generated under this are of research.

Reflected to the Table 3 above, it can be concluded that the content of NEGEM database largely fall under types of model-based simulation data and, on the other hand, observational data from surveys. Noteworthy,

the EC open access manual[13] largely defines research data under open access initiatives as 'underlying data' (such as the data needed to validate the results presented in scientific publications). Still, inclusion of public NEGEM deliverables and other documents in the NEGEM database can be considered to support findability and exploitation of the results.

3.2   *Identified restrictions for NEGEM data*

As identified in the DoA, the database development needs to comply with the GDPR, the General Data Protection Directive (2016/679) of the European Parliament and of the Council. Accordingly, data privacy will be respected and carefully monitored. To strictly comply with the directive, an approach is taken that any data put available in the NEGEM Open Access database need to be anonymous. Especially, this is relevant for the organisations managing NEGEM data who apply the methods of social sciences, that in many cases need to manage personal information of the interviewees etc. Hence, in every research phase, ethical procedures and legislation needed to comply with this requirement are followed in these activities. As an example, the conflict between data privacy and open access was elaborated in the case of feasibility of storing pseudonymized transcriptions of interviews in database bult up for earlier H2020 project (ECHOES). There, after consultations with experts, a decision was made to rule this material out of the Open Access database, even if pseudonymized, due to potential endangerment of the privacy of interviewees.

Technically, there are also very data-intensive methods applied in NEGEM. Hence, as there are size limitations for open access research repositories such as 50 Gb / dataset, a demand for tailored solution for large datasets is identified.

As third type of restrictions, many datasets relevant to NEGEM are subject to copyrights or fee or other contractual IPR obligations. Hence, all the data potential to be stored in NEGEM database will be checked for these issues before stored or made public.

---

[13] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

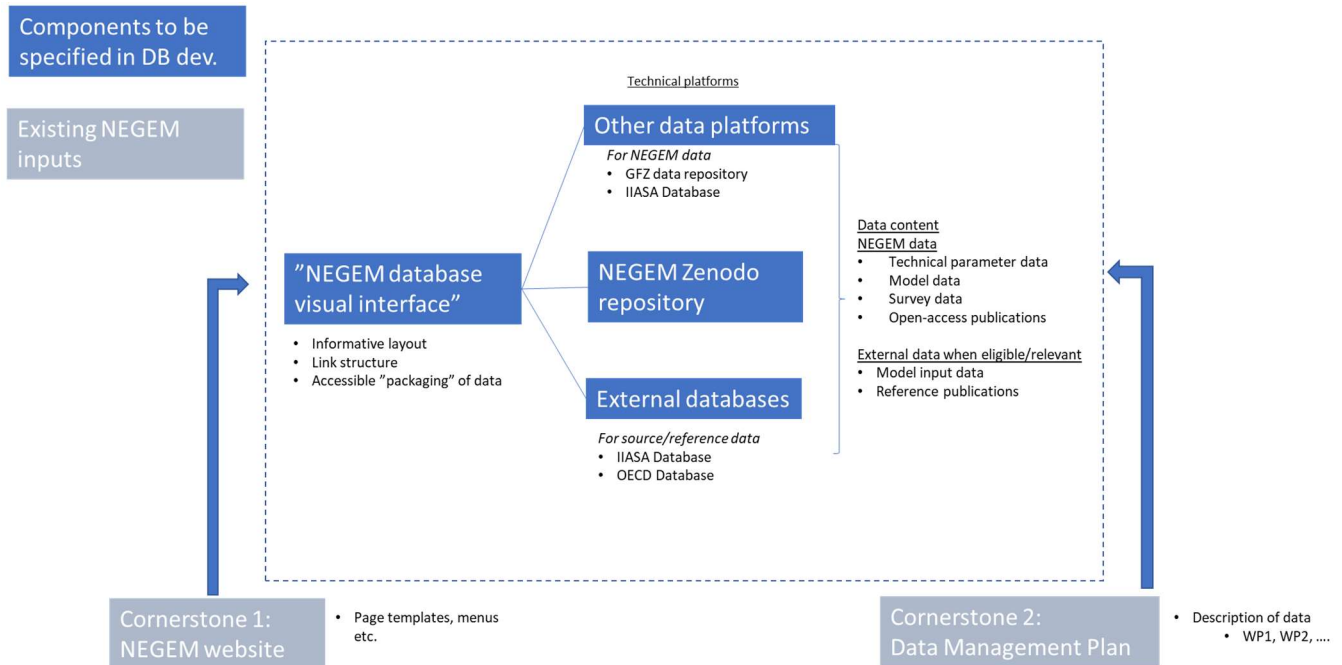## 4 NEGEM database components

Figure 1. NEGEM database components

Figure 1 describes the components of NEGEM database from technical point of view. Components of NEGEM database include linkages to external databases (DBs), as well as tailored data repository for NEGEM data. The components are differentiated between those available in the first period of the project (Month 1 – Month 18) (Cornerstones), and the components under development. These components will be implemented based on the plans presented in this deliverable. Different outcomes of the NEGEM project Work Packages for the database will be completed at different times during the four-year-project. This implies stepwise addition of data in the database, depending on detailed work plans and publication plans of different Work Packages. During the active project phase, coordinator and in partner in charge of task 8.6, VTT, will manage the database. In any update needs, e.g. due to expiring of links referred, partners can contact VTT to correct the information in collaboration with partner ETA managing the NEGEM website. Also, records can be updated in major needs even after the project. This can have importance during the first few years after the project, when data collected and generated can still be considered mostly up-to-date.

From the end-user point of view, the NEGEM database will be built on the visual interface to be implemented in the NEGEM website (*Cornerstone 1, lower left-hand-side corner in the Figure 1*). The service on the website acts an entry point for the access of data. From the data content point of view, Data Management Plan (DMP, D10.3, D10.5, D10.7, *Cornerstone 2, lower right-hand-side corner in the Figure 1*), including a thorough description of the data generated in different phases and sub-tasks of the NEGEM project, is utilised as a key source. However, the Data Management Plan does not focus on the viewpoint of making the highly multi-disciplinary data of NEGEM accessible. Especially, scattered data stored in various, potentially domain-specific platforms and repositories do not form a consistent legacy for the project after it is completed. Hence, from this viewpoint, the task for a NEGEM database development is to collect and organize the relevant NEGEM data as described in Data Management Plan in its entirety to a user-friendly service.

In NEGEM open access database development, scientific exploitation of the data has been identified as an important target due to the nature of the project.   Also, the restrictions identified in data management (e.g. very large sizes of data limiting the use of repositories, privacy issues and other ethical reasons limiting making data public, etc.) make it necessary to determine the best portfolio for different data repositories and databases for NEGEM data. From these viewpoints, the core task in NEGEM database development is to determine eligible data for open access and collect and structure it logically and accessibly, considering the multi-disciplinarity of the project and meeting the appropriate scientific standards.

In the following sub-sections, the components of the NEGEM database are discussed further, to address the above-mentioned principles and to present a variety of options for the database. The exact solutions applied are more thoroughly determined in the implementation phase of the database (M20-M48) in collaboration with NEGEM partners representing different disciplines. Stepwise approach allows to update the initial plans presented below with any aspects that may appear with the data as the project results and other NEGEM learnings accumulate.
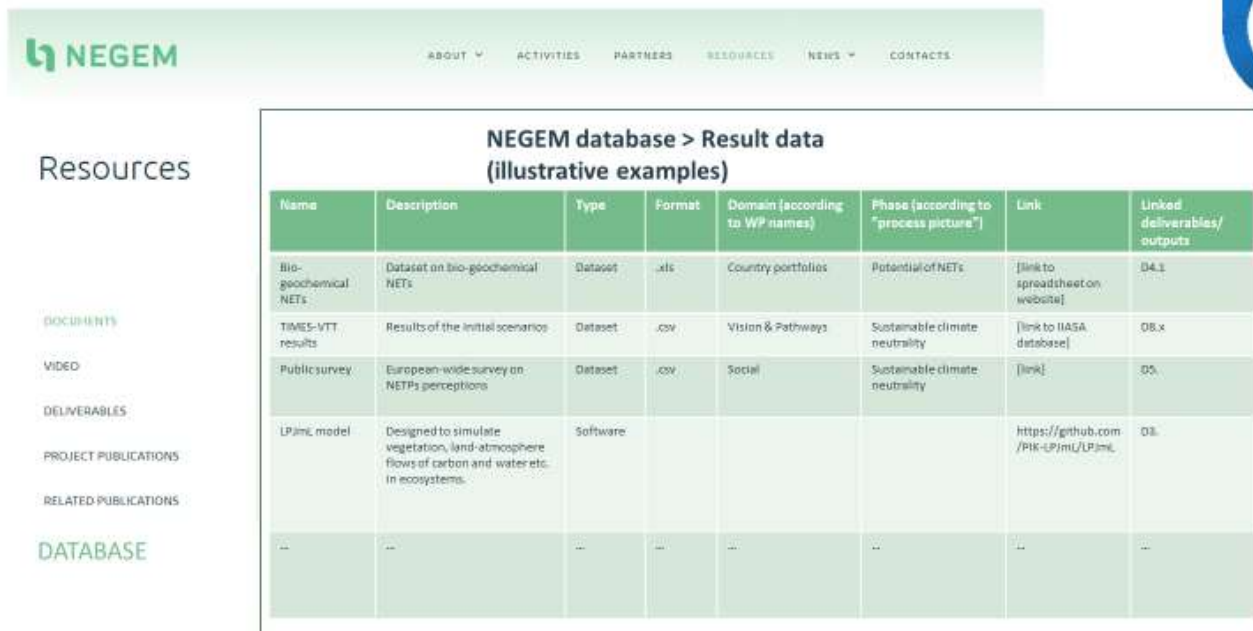
4.1   User interface

User interface of NEGEM database is developed on the website of NEGEM (www.negemproject.eu). The user interface is designed in co-operation between VTT and ETA as a partner responsible of website development and dissemination of the project. Examples for potential functionalities for the NEGEM database can be explored in database available in ETIP Bioenergy website[14] implemented and managed by ETA. As key design requirements for the NEGEM database, easy navigation and access for any data put available, and allowing simple exploring of data to an extent, have been identified. Accordingly, the "landing page" for the NEGEM database allows the user to explore the following options as key content if relevant:

- NEGEM result data
- Source data
- NEGEM data repository

After choosing the area of interest, filtering of data records, e.g. according to type of data, phase of the research and domain of the data, is possible for the end-user. Forthcoming data records can be labelled with "To be specified" or "To be added" when relevant. Thereby, clear, and extensive list of already published and future NEGEM data records can be explored already during the project by interested end-users. Figure 2 demonstrates the structure of the website and functionality of exploring *NEGEM result data* available.

---

[14] https://etipbioenergy.eu/databases/reports

Figure 2. Sketch of the NEGEM database user interface to be implemented in the NEGEM website.

Especially, the following features are implemented for browsing the "*NEGEM result data*" records:

- Option to show all the data records at once, with possibility to filter by the following attributes when relevant:
  - Name
  - Description
  - Type: Dataset, software, deliverable, etc.
  - Format[15]
  - Domain
  - Link to access the data (labels "To be specified" / "To be added" applied when relevant)
  - Linked Deliverable / output

The following features are implemented to serve the goal of usability of the database and its high exploitability.

- Records can be flexibly added, also for intermediate data. The exact procedure will be agreed and communicated in the consortium
- The user does not need to be aware of internal project management such as WP structures but can browse the material without any prior knowledge

---

[15] Regarding formats of the modelling files, original files can be kept, and whenever possible, also general format files are provided. In optimization, for example, GAMS file and also the model in the general MPS format, could be provided in the database.

The "*Source data*" option can be implemented provided there are sufficiently records to justify it as a separate category. It can be implemented straightforwardly as a thematically organised link list. Alternatively, a table structure similar to "NEGEM result data" records described above (Figure 2), can be used.

The "*NEGEM data repository*" option provides link to external research data repository intended for permanent storing of NEGEM data. Especially, this task is important as there are no resources allocated to update the NEGEM website constantly, but such limitation does not to apply to research data repositories designed for permanent storing meeting the scientific standards. Importantly, including tailored repository as a NEGEM database component enables DOI numbers, search functions, permanent storing and similar functionalities important and required for scientific open access purposes. On the other hand, structural options to present data are relatively fixed in ready-made research repositories. This motivates to develop a user interface for the NEGEM database that flexibly allows structures such as folders or menus.   The user interface will be made easily accessible through the NEGEM website. The NEGEM website has been built on Wordpress platform and there are suitable add-ons for the desired functionalities. Implementation of the database is seen most fluent as static solution.

4.2   Repositories for NEGEM data

4.2.1   Background

There are a plentiful of commercial and open repositories available for research data. They can further be divided in generic data repositories and discipline-based repositories.

Based on internal scanning work at VTT for the ECHOES project with a reliance on expert on knowledge services, characteristics of some EU based, international and national repositories were studied. Accordingly, there are at least the following three alternatives for choosing the generic repository for NEGEM as another H2020 project were identified:

- *Zenodo*[16] funded by CERN and EU OpenAIRE project[17]
- *EUDAT[18]*, a research data infrastructure initiative funded by European Commission
  - *B2SHARE*: simple solution for small-scale sharing of research data
  - *B2SAFE*: robust and safe replication service for local repositories with PIDs and metadata in place
- *Figshare[19]* by UK based Digital Science & Research Ltd: offers easy-to-use data depositing and publishing solution

The data repositories are not limited to three options mentioned above. Wilkinson et al. (2016) mentions *Dataverse* (https://dataverse.harvard.edu/), *Mendeley Data* (https://data.mendeley.com/), *DataHub* (http://datahub.io), and *DANS* (http://www.dans.knaw.nl/) as additional examples of open globally-scoped repositories.

---

[16] https://zenodo.org/
[17] https://www.openaire.eu/
[18] https://www.eudat.eu/
[19] https://figshare.com/

In addition to generic data repositories, web-based search service *re3data* demonstrates the large variety of available discipline-based repositories[20].

Due to multidisciplined nature of NEGEM, the flexibility of generic data repositories justifies their consideration for a centrepiece for the NEGEM database. Additionally, the suggested structure of NEGEM database allows the use of discipline-based repositories when needed. The user interface available on the NEGEM website and/or the central NEGEM repository would be used to guide the user in such cases.

4.2.2   Zenodo repository

Zenodo repository was recommended by knowledge services personnel of VTT and introduced as ECHOES project data repository. Application of Zenodo was justified by the fact that Zenodo was mentioned in the Guidelines for FAIR Data Management in Horizon 2020 and among the EU OpenAIRE recommendations for finding a research data repository.

Due to the above reasons, Zenodo repository was introduced in the NEGEM project plan and it is further elaborated in this plan as a "default solution" for creating central NEGEM repository. There is a reason to believe that requirements on permanence of storing, reliability, and security can be efficiently met through applying the alternatives referred in H2020 materials for open data.

Zenodo repository has its qualities and limitations[21]. One of its important dimensions is meeting the scientific requirements on data to be stored such as metadata definitions. On the other hand, the structures are relatively fixed, and limitation for size of the data records may constraint storing of NEGEM data. Especially, the following characteristics are noteworthy:

- Zenodo provides a permanent identifier for research data (DOI)
- Zenodo enables storing of different type of materials, including publications, posters, presentations, research data, images, videos, audio files, software and lectures
- *Communities* feature offers an option to form a group over a certain topic and to curate data to be published
- As a default, size of data record allowed limited to 50 Gb
- Can be integrated to GitHub, allowing references to source codes / softwares via DOI
- It is possible to deposit only metadata in Zenodo
- Publishing of versions of research data supported

Regarding the Zenodo implementation for NEGEM, a creation of NEGEM community is planned. The community would include a reasonable number of records, functioning as a permanent storage for NEGEM results and fostering their use. Each of the records would include collections of key files of the part of NEGEM in question. Also, links to external storages could be included if necessary (e.g. due to size restrictions). Particularly, the following aspects are highlighted features in the records to be available in NEGEM community.
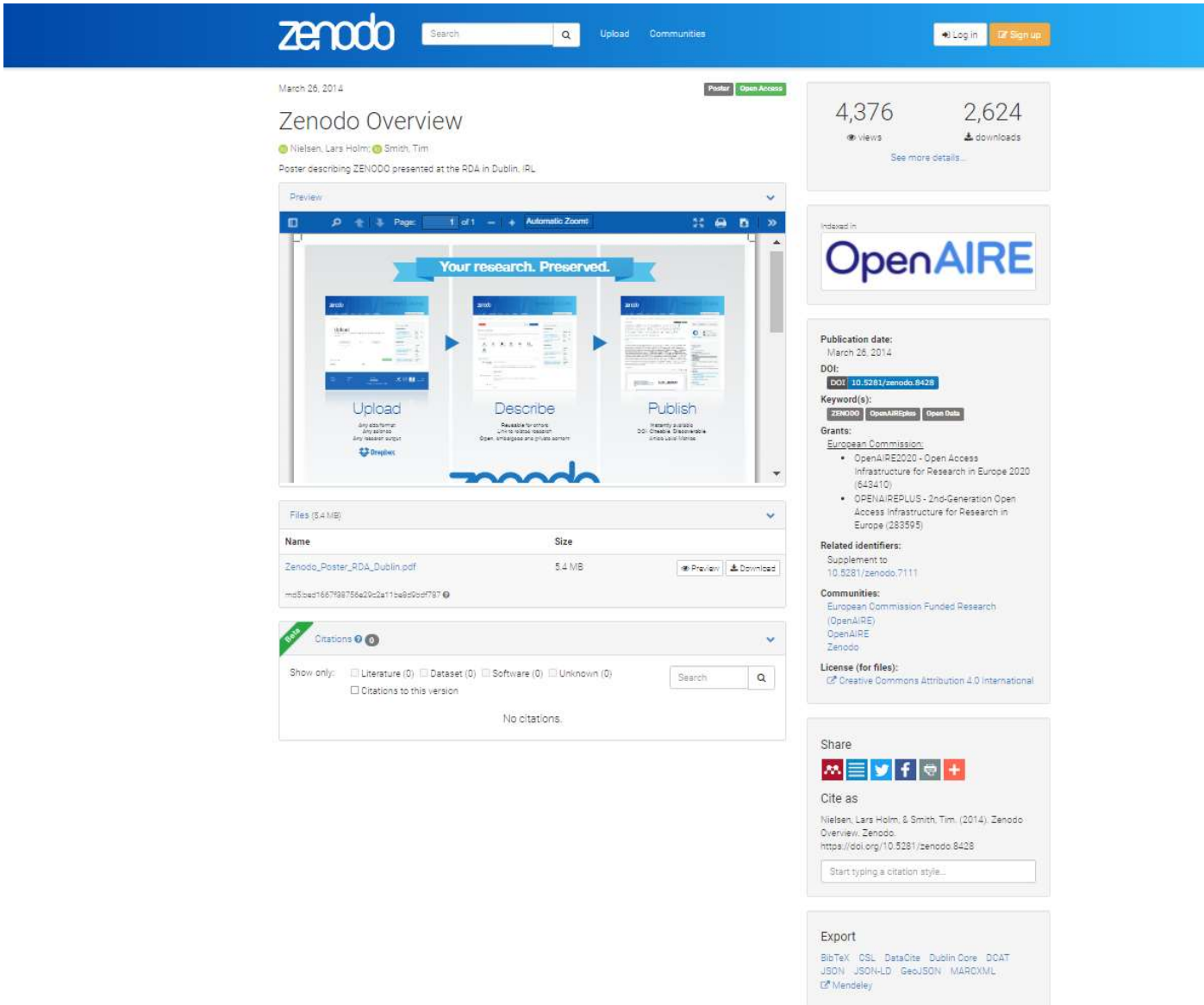
- DOI numbers and similar scientific requirements are automatically created in line with OpenAIRE, FAIR principles etc.

---

[20] http://www.re3data.org/
[21] https://www.aalto.fi/fi/palvelut/zenodo-data-arkisto

NEGEM

- Additional information to be included for records is designed for scientific purposes, meeting standards on metadata, etc. according to H2020 guidelines

- As new versions can be added but published record cannot be changed, the repository suits better for "more mature" data than very preliminary intermediate work versions

Figure 3 demonstrates features of a data record stored in Zenodo repository. As can be seen, the record includes typical scientific information such as authors, type of data, descriptions, recommended citations etc., list of files included in the record and license applied for the record. Searches by keywords, names, and other typical functionalities for research repositories are possible.

Figure 3. An overview of the Zenodo record and available information (source: https://zenodo.org/record/8428#.YcLimGBBxaQ)

4.2.3 Parallel repositories

Application of Zenodo repository does not rule out application of parallel, e.g. discipline-based repositories in the NEGEM database if needed. These can be linked and/or introduced through the user interface implemented in the NEGEM website. Regarding NEGEM data, it has been identified in Data Management Plan, for example, that partner PIK applies models that can be very data intensive, resulting in huge requirements for data sizes, exceeding the limits of Zenodo repository. Hence, as an example of tailored solution, the GFZ repository[22] is applied for this data. Also, research (reference) repositories of partner organisations that are systematically used to archive outputs of the partners, e.g. Oxford University Research Archive[23], represent parallel repositories for the project outputs, and also they can flexibly be linked to the central NEGEM database records when needed.

4.3 Linkages to other databases / sources

To guarantee accessibility of data, NEGEM utilizes ready-made widely known databases whenever relevant. Noteworthy, careful consideration for a proper presentation of data is needed in cases where external databases are not open source (e.g. Ecoinvent). External databases can be relevant for source data and, on the other hand, as platforms for NEGEM data. In these cases, the task of the NEGEM database development becomes presenting the data stored in external databases in easily accessible layout. *Table 4* presents the external data platforms that have been identified relevant, and their inclusion to the NEGEM database will be considered in further deployment work. Table 4 may be supplemented by alternative sources as the project proceeds. For example, data attached to Open Access publications in databases of different scientific publicators may provide alternatives additional to mentioned in Table 4.

Table 4.External databases used as input or output for the NEGEM project. Contents of different Work Packages: WP1: Technology assessment, WP3: Impact Assessment, WP8: Vision and framework for pathways analysis (Source: modified from DoA)

Database/Owner	Work Packages	Expected outcome
TIMES-VTT database/VTT	WP3, WP8	NEGEM Pathways and impact assessment
IEAGHG and energy statistics/ IEA	WP3, WP8	NEGEM Pathways and impact assessment
SSP/IIASA and others	WP3, WP8	New information on NETPs and its role in 1.5 °C pathways[24]
Agri-Footprint/Blonk Consultants	WP1, WP3	Life Cycle Impact modelling
Ecoinvent/ETH and Agroscope Life Cycle Impact modelling	WP1, WP3	Life Cycle Impact modelling

---

[22] https://dataservices.gfz-potsdam.de/portal/

[23] https://ora.ox.ac.uk/objects/uuid:3eebf969-1661-4484-a510-89470602cada

24 IIASA's database collects documentation of quantitative projections of the Shared Socioeconomic Pathways (SSPs) and related Integrated Assessment scenarios. See https://secure.iiasa.ac.at/web-apps/ene/SspDb/dsd?Action=htmlpage&page=about

5   *Key findings and policy relevant messages*

This deliverable presents key building blocks and design premises for open access NEGEM database, presenting a "legacy" of the project and aiming at high utilization of project data already during it. Building of the database is seen, especially, to support scientific exploitation of the project's non-commercial outcomes. Thereby, the process has potential to contributes to science-based policymaking on primarily European, but also on global and national levels in the areas of negative emission technologies and practices.

Based on the reasons of distributed management and execution of data collection efforts, to enable flexibility, the proposed approach for NEGEM database can be called a modular "collection" of databases, datasets and other sources.

NEGEM database is open to any stakeholder group or private person interested in the results. As the results also aim at high policy relevance, the exploitation groups connected to policymaking are even more directly targeted with other tailored outputs such as policy briefs, newsletters, targeted deliverables, and dissemination events. To foster the exploitation of the data, these outputs are put easily accessible and findable via the same user web-based interface as applied for the database.

While open data in general can create unforeseen paths, synergies, and new innovations for utilization of data, still, especially, the following data primarily landed in NEGEM database are seen to have potential in supporting policymaking related to NETPs:

- Data on technical and cost parameters, such as potentials and constraints as well as techno-economic, environmental, or other characteristics of NETPs, contributing to the relevancy and consistency of modeling efforts to determine sustainable pathways in meeting the Paris Agreement goals;
- Data on visions of role of NETPs in climate dimension in global, European, and Member State levels for creating roadmaps and scenarios supporting in developing climate change mitigation strategies;
- Data on perceptions of different stakeholder groups based on surveys, helping to determine efficient strategies to improving acceptance of sustainable NETPs.

6 Conclusions and further steps

6.1 Conclusions on the NEGEM database design

Building up a database including NEGEM results and relevant external data calls for a careful plan, and there are several building blocks to be decided and several project partners involved. Up to a point of release of this deliverable (D8.5) of the NEGEM project (January 2022), the following steps have been taken:

- Mapping the data content of NEGEM preliminarily suitable for the database, based on preparing Data Management Plans of NEGEM (all NEGEM partners)
- Mapping of different technological options and structures needed to implement the database (VTT development work in collaboration with ETA on user interface and related structures on NEGEM website)

Based on these steps, the following key design guidelines for the building blocks of the NEGEM database are concluded:

- **User interface**
    - To be integrated on NEGEM website with collaboration of VTT and ETA
    - Primary use case will be based on scientific exploitation
- **Data content**
    - Meeting the Grant Agreement criteria on reproducibility of scientific publications
    - Further domain-specific dialogue on any relevant additional content relying on structure of the database
    - No confidential or private data will be stored in the central NEGEM database
- **Technical solutions**
    - Repository for long-term scientific archiving (e.g. Zenodo recommended by the knowledge services of VTT)
    - Flexible solution to be implemented on NEGEM website for curating intermediate data and making it findable and accessible
    - The NEGEM website has been built on Wordpress platform and there are suitable add-ons for the desired functionalities. Implementation of the database is seen most fluent as static solution. The update procedures will be agreed with ETA and rest of the consortium and can be built e.g. on tailored Excel worksheet or informing ETA whenever there are update needs.
    - Security is ensured with recommended measures and software by organisations implementing the database, including use of authentication methods such as passwords.
    - Long-term and secure preservation of published research data will be ensured by using only certified and OpenAIRE guidelines compatible repositories.

Considering the development of NEGEM database, a table or link list implemented in the NEGEM website appears to be the most convenient and flexible solution for the inclusion of the various types of information in organised form. In this case, characteristics of database such as searchability, are on one hand, implemented in the user interface and, on the other hand, available through services available on research-based data platforms.

All encompassingly, there is a need to pay attention to choices and their systematic collection such as metadata definitions, user interface design, and accessible categorising, in order to make the database eligible from the FAIR data management principles point of view. This report identified several technical alternatives as the building blocks of the data management infrastructure that will be further developed to meet the project objectives on exploitability in the upcoming work.

6.2 Further implementation

In NEGEM, management and execution of data collection efforts are multi-disciplinary and distributed between the partners. To enable flexibility, NEGEM database can be called a modular "collection" of databases, datasets and other sources. For efficiency benefits, practical fluency, and partners' established conventions regarding data, the NEGEM database solution highly relies on application of available tools rather than building an all-encompassing solution from the scratch. To implement the database, a partner responsible for task 8.6, VTT, will run the implementation process as follows:

1. Stepwise implementation of components of the database in dialogue with partners based on this plan, starting after agreeing on it and communicating with the partners (1/2022).
   - Leveraging established data platforms and conventions of the partners, thus avoiding unnecessary and overlapping labour-intensive development work.
2. Structures of the NEGEM database will be completed in collaboration with ETA and VTT early 2022.
3. Set-up of structures enables step-wise addition of the data, further demonstrating the database during the second reporting period (ending November 2022). Building on this, the database completely up and running by the project end (May 2024).
4. When finished, the NEGEM database will be supplied with all the relevant NEGEM data, enabling its permanent storing in line with H2020 requirements on open access data.

For preparing this report, the following deliverable/s have been taken into consideration:

D#	Deliverable title	Lead Beneficiary	Type	Dissemination level	Due date (in MM)
D4.1	NETP database	ICL	Other	PU	M9
D4.2	Bio-geophysics database	ICL	Other	PU	M12
D10.3& D10.5	Data Management Plan	VTT, collaboration with all partners	ORPD	CO	M6, M18
D11.1	POPD-requirement no 2	VTT	Ethics	CO	M3

References

Correia, S., Similä, L., Piira, K., Kannari, L., & Koljonen, T. (2018). H2020 ECHOES Deliverable D2.2: ECHOES database description. European Commission EC. https://echoes-project.eu/sites/echoes.drupal.pulsartecnalia.com/files/D2.2.pdf

EC, European Commission, Directorate-General for Research and Innovation (2021). *Horizon Europe, open science : early knowledge and data sharing, and open collaboration*, Publications Office, 2021, https://data.europa.eu/doi/10.2777/79699

EC (2020). Horizon 2020 projects working on the 2019 coronavirus disease (COVID-19), the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and related topics: Guidelines for open access to publications, data and other research outputs. Version 1.1 April 24, 2020. https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-guidelines-oa-covid-19_en.pdf

Similä, L., & Koljonen, T. (2017). *Report of existing data and databases as additional to new data collection and formulation: H2020 ECHOES Deliverable D2.1*. https://echoes-project.eu/sites/echoes.drupal.pulsartecnalia.com/files/D2.1.pdf

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). *The FAIR Guiding Principles for scientific data management and stewardship.* Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18