Nordic Council of Ministers

Nordic cooperation on data to boost the development of solutions with artificial intelligence

Contents

Preface

Data and artificial intelligence play an increasing role in our daily lives and hold a great potential to further improve our society. By leveraging the enormous amount of data generated by public authorities every day, we can improve efficiency of both the public and private sector. Better access to data combined with the responsible use as well as the opportunities offered by artificial intelligence could, for instance, enable our health care system to save more lives, make our businesses more cost-efficient and help us combat climate change more effectively. All to the benefit of society.

Together, the Nordic countries manage a huge amount of government owned data. By making more of the data easily available to both public and private entities, the Nordic countries can boost the development of better digital services and solutions, thus contributing to innovation and growth of society. This report serves as a contribution to enhance Nordic cooperation on data access with the aim of boosting development of new and innovative solutions with artificial intelligence.

The report is the result of a Nordic study conducted by Rambøll Management Consulting. Based on an assessment of different government owned datasets, the report provides recommendations of how to overcome barriers for more efficient data sharing. This report constitutes a first step in the identification of government owned datasets across the Nordics that has artificial intelligence potential.

The recommendations in the report will be discussed at the Ministerial meeting for Nordic Business Policy on September 1st, 2020 and the report will also serve as a key input to the joint Nordic cooperation on AI and access to data which with respect to the joint Nordic Action plan for the Nordic Council of Ministers Vision 2030 during 2021-24.

Executive summary

This study presents an initial overview of potentially relevant government owned datasets in and across the Nordic countries with high value for developing artificial intelligence (AI) solutions while identifying barriers related to openness and AI usage and providing recommendations on how to address these barriers.

The study has focused on formulating policy recommendations for breaking down barriers and enhancing knowledge on and availability of government owned datasets for AI solutions. Despite emerging political focus in the Nordic countries on the need to prioritize AI development to improve efficiency, deliver better quality services in the public sector and ensure competitiveness in the private sector, AI initiatives remain fragmented and investment appears low. The recommendations aim to challenge these issues by identifying appropriate joint Nordic actions which can act as a solid starting point for the Nordic Council of Ministers' working group on AI and Access to Data.

The outcome of the study has been reached firstly by establishing a set of criteria for valuable datasets to boost the development of AI solutions, thus improving competitiveness of businesses and increasing societal value. Secondly, identifying an initial list of datasets by means of a detailed desk study and interviews with owners of government datasets. These datasets are assessed to be of potential high value for AI development and for businesses. The datasets have been ranked in order to identify which has the greatest potential value and highest relevance for AI solutions.

The report is intended for policy makers aiming to further the open data agenda and growth of AI business sectors, and data owners in public organizations seeking inspiration on how to address and overcome the barriers related to making their datasets available to the public. Technical details and considerations have been relegated to the appendixes and the report is kept short and accessible for non-experts.

Identified top-ranking datasets

The top-ranking datasets identified include Groundwater, Weather Data, Road Cameras (Photos), Traffic Events and Roadworks, Energy Data (Aggregates) and Company Announcements. All of them have been classified as having a very high value for developing AI solutions in the Nordic countries.

Common for the datasets is that they have a high or very high **estimated value for businesses and low or moderate barriers for value-realization**. They are characterized by having no obvious legal barriers averting the process towards making the dataset available to the public, e.g. by not containing sensitive information. Regarding the quality of data and the work required to make the datasets ready for publication and to maintain them once they are made publicly available, the datasets have assumed low costs associated with them. Also, for the datasets, there is a clear ownership and a responsible organization.

Only two of the datasets on the list have been assigned a score of Very High **cross-Nordic value**. This is the Groundwater data and the Weather data. Both datasets are very similar across the Nordic countries with respect to variables and information and characterized by a high degree of openness of data across the Nordic countries, enabling possibilities of strong Nordic collaboration.

The datasets with a high or very high **societal value** are datasets considered useful in AI applications or AI solutions for a greater good, such as having a positive impact on climate change and improving the health and well-being of Nordic citizens. This is applicable for e.g. Weather data. Weather data is vital for both business and society. Weather forecasts are important for e.g. agriculture and road transport, and rain forecasts play an important role in e.g. wastewater treatment.

Finally, the datasets have also shown a high or very high **AI-relevance** indicating that they are available in AI relevant formats, in the context of size, richness and that they contain labels. Some of the highest scoring datasets are those collected by sensors or similar machine-operated devices. This includes Groundwater data, Weather data, Energy data, Air quality data and the Road Cameras data.

The project concludes that **most datasets can be found as open datasets** in at least one and often several of the Nordic countries. This provides ample opportunities in all data domains for strong Nordic collaboration.

Recommendations

Overall, the study concludes that government owned datasets can be made more visible to companies, generating interest in datasets and potentially a business demand for making datasets publicly available. This can help the public sector identify which datasets to make public first. Visibility can be furthered through hackathons, promoting the Nordic open data portals and encouraging public organizations to publish information on datasets that have yet to be made publicly available.

Furthermore, these datasets can be made more available for AI solutions and development. This can be done by providing easier access through e.g. APIs, releasing metadata and dataset descriptions alongside the datasets, and providing dataset information in more than the local language.

The following joint Nordic actions are suggested as next steps for furthering the open data agenda in the Nordic countries and creating opportunities for companies wanting to create AI solutions on government owned datasets:

1. Arrange cross-Nordic hackathons on government owned data
2. Collect and showcase examples of the value of government owned data from across the Nordic countries
3. Fund projects creating an overview of which government owned datasets are highly used and demanded by companies across the Nordic countries
4. Establish Nordic working group on open data standards and formats including best practices when publishing data
5. Fund projects investigating the potential of new or known methods to publish sensitive data in accordance with GDPR
6. Fund projects to make groundwater data and road camera data more

accessible for companies across the Nordic countries
7. Promote the open data portals in the Nordic countries
8. Collect good practice examples from the Nordic countries on good data governance and data management related to publishing datasets

The joint Nordic actions presented above stem from 10 main barriers identified in the project that are relevant for four different sets of stakeholders. For each barrier, one or more recommendations have been identified to help overcome it. The barriers and recommendations are presented in detail in chapter 3 of the report.

*Recommendations on how to address barriers relevant to the **National level**:*

1. Collect or construct commendable showcases and examples of the value of government owned open data from across the Nordic countries
2. Use data format recommendations and standards; in particular international standards when available
3. Facilitate emergence of data ecosystems
4. Exemplify needs and avoid building proprietary solutions, whenever possible
5. Find ways of funding to compensate public organizations that are publishing data at a cost for businesses
6. Enlist the help of citizens, startups and the open data community
7. Encourage and support collaboration with startups and SMEs
8. Fund projects investigating fully GDPR compliant options for releasing sensitive information
9. Encourage public organizations to release sensitive datasets in an aggregated form

*Recommendations on how to address barriers relevant to **Data generators**:*

1. Ensure that data management and the data architecture promote easy overview of and access to data
2. Enable data re-users to help improve dataset quality
3. Make datasets available with documentation of the processes that were used to create a specific dataset

*Recommendations on how to address barriers relevant to **Data publishers**:*

1. Facilitate making data publicly available through a standardized data submission setup
2. Encourage data generators to utilize professional data publishers

*Recommendations on how to address barriers relevant to **Data re-users**:*

1. Create visibility for the datasets that can be made publicly available
2. Promote the open data portals in the Nordic countries
3. Undertake preliminary work into the creation of a cross-Nordic open data portal
4. Communicate the purpose for which the datasets have been collected
5. Engage in public-private dialogues with the market

6. Publish datasets with proper and detailed data descriptions
7. Provide guidance for data publishers on the European metadata specification DCAT-AP

Finally, the project has gone into detail with two of the datasets with the highest assessed value: Groundwater data and Road camera data (photos). The report contains specific recommendations on which actions need to be taken by which public organizations in which countries with respect to making these datasets more accessible for companies. Weather data is assessed to be of higher value than Road camera data (photos) but work is already underway in Denmark in making this data publicly available and Weather data has thus not been selected for further inquiry.

The report has been produced by Rambøll Management Consulting in collaboration with Research Institutes of Sweden, on behalf of the Nordic Council of Ministers. We would like to thank agencies and stakeholder that have provided input and information throughout the study.

Introduction and background

The Nordic countries, individually and especially jointly, have a solid pool of public data that could be used to create value for businesses and society when made publicly available. Strategic government work has been done in all the Nordic countries, setting national aims for the use of AI and identifying barriers to overcome in order to reap the benefits.

The development of AI solutions has been heralded as disruptive change to almost all parts of the economy[1] [2] [3], including the public sector[4]. There is no longer any doubt that AI has a very considerable potential in developing and implementing AI solutions for addressing environmental and social challenges in society at large[5].

However, access to data is of crucial significance for the development of AI solutions[6], and despite all efforts and investments so far, governments in the Nordic countries are still encountering barriers in making data publicly available and companies are facing major obstacles finding, accessing and re-using government datasets and sources. These are the key issues addressed in this report.

The purpose of this report is to establish criteria for which datasets that are most valuable to the development of AI solutions, thus improving competitiveness of businesses and increasing societal value. Furthermore, the report will provide an initial overview of potentially relevant government owned datasets in and across the Nordic countries with high value for developing AI solutions while identifying barriers related to openness and AI usage and providing recommendations on how to address these barriers.

The report is intended for policy makers aiming to further the open data agenda and growth of AI business sectors and for data owners in public organizations seeking inspiration on how to address and overcome the barriers related to making their datasets available to the public. As such, technical details and considerations have been relegated to a rich set of appendixes while the report is kept short and accessible for non-experts.

Open Data and AI in the Nordic Countries

Government owned data can in the private sector be used as digital raw material for developing digital services and digital content, thereby contributing to innovation and growth. Other sectors can use owned data to create new intelligent services, advanced analyses and targeted information for the benefit of both citizens and companies. In this way, new digital markets can be created, and government owned data can contribute to innovation and growth.

1. An AI-nation – Harnessing the opportunity of AI in Denmark (The Innovation Fund Denmark and McKinsey & Company, 2019) & Nordic municipalities' work with artificial intelligence (Ulf Andreasson and Truls Stende, 2019),
2. Artificial Intelligence in Swedish Business and Society (Vinnova, 2018)
3. How Artificial Intelligence Will Transform Nordic Businesses (McKinsey & Company, 2019),
4. Främja den offentliga förvaltningens förmåga att använda AI (Myndigheten för digital förvaltning (DIGG), 2019)
5. Artificial Intelligence in Swedish Business and Society (Vinnova, 2018)
6. Artificial Intelligence in Swedish Business and Society (Vinnova, 2018)

The Nordic countries are working together to ensure that the Nordic region remains a digital frontrunner. The countries have agreed to cooperate closely on the topic of AI, resulting in a "Declaration on AI in the Nordic-Baltic Region" (May 2018) by the ministers responsible for digital development from Denmark, Estonia, Finland, the Faroe Islands, Iceland, Latvia, Lithuania, Norway, Sweden and the Åland Islands.

The Nordic countries are generally very open with regard to public data across most data domains according to organizations such as the European Data Portal[7], Open Data Watch[8], Open Knowledge Foundation[9] and the World Wide Web Foundation[10]. The Nordic countries offer open government data to the public through dedicated Open Data Portals, such as data.norge.no in Norway and avoindata.fi in Finland.

AI plays one of the most important roles in the data economy, and access to open datasets is a crucial ingredient to achieve the potential of AI to "*help solve major societal challenges and provide significant benefits in a variety of areas*", quoting the previously mentioned declaration on collaboration on AI in the Nordic-Baltic region. AI and machine learning algorithms are used to extract general insights from large amounts of data. Since these algorithms can only learn from what is in the data, open datasets for AI need to be of good, controlled quality and contain substantial, varied and trustworthy information[11]. Public governmental datasets are good candidates for open data since they generally already fulfil these criteria[12].

In the European Union (EU), legislation is continuously being adopted to foster the re-use of open government data in the member states. In 2015, a report procured by the European Commission estimated that the market size of open data was expected to increase by 36.9% from 2016 to 2020, to a value of 75.7 billion EUR in 2020[13].

Recent legislation adopted in the EU is the recasted Public Sector Information Directive (the Open Data Directive, ODD), which among other things calls for "*the provision of real-time access to dynamic data via adequate technical means, the increase of the supply of valuable public data for re-use*"[14].

The EU runs its own Open Data Portal[15] and have pushed persistently for publishing and pooling datasets from across the EU member states, e.g. by facilitating public access to spatial information across Europe through the INSPIRE Directive[16]. The ODD also introduced the concept of High Value Datasets (HVD), defined as "*documents the re-use of which is associated with important benefits for the society and economy*". The EU member states have been given the task to supply their national HVDs, intended for inclusion in the European Open Data Portal. The Nordic HVDs have provided valuable input to this report.

7. https://www.europeandataportal.eu/en/impact-studies/country-insights
8. https://opendatawatch.com/
9. https://index.okfn.org/
10. https://opendatabarometer.org/barometer/
11. Declaration on AI in the Nordic-Baltic region (2018; https://www.norden.org/da/node/5059
12. AI and Open Data: a crucial combination (2018; https://www.europeandataportal.eu/en/highlights/ai-and-open-data-crucial-combination
13. Creating value through open data (Carrera, Chan, Fischer, & van Steenbergen, 2015)
14. Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information
15. https://data.europa.eu/euodp/en/home
16. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)

Assessing the value of government owned datasets

To assess the value of a government dataset that has not yet been made publicly available, in terms of the potential for the development of AI solutions, a framework consisting of a set of criteria has been developed, synthesized from similar projects on the value of Open Data[17] [18]. The framework and the full list of criteria is presented in Appendix 1. In brief, the framework employed to assess and score the datasets in this report consists of five elements. Each of these elements will be briefly described in the following sections.

1. AI-relevance
2. Barriers to value-realization
3. Societal value
4. Estimated value for businesses
5. Cross-Nordic value

Since most AI applications require data of certain volume, the criteria developed to measure AI-relevance has been inspired by the literature on big data, where big data often is described in terms of the four V's: Volume, Variety, Velocity and Veracity[19] Moreover, to develop valuable AI solutions, datasets are more relevant if they contain labels or similar that can be used for prediction and classification.

With regard to barriers for value realization, in general two perspectives are predominant. Firstly, there are barriers for governments and governmental agencies related to legal issues, costs and technical competencies. Secondly, there are barriers for the data users in terms of data access and the quality of data provided. Both perspectives are included in the assessment of datasets.

In this report, societal value of a dataset is evaluated as the potential of dataset supporting achievement of societal goals. In the Open (Government) Data literature, more open data on government operations are believed to improve the quality of a democracy in a country, simply by letting non-governmental agents in a country monitor what the government is doing and how[20]. The approach pertaining to societal goals in this framework also builds on this assumption, assuming that (more) data on e.g. air quality can help citizens and companies monitor what the government is doing to reduce air pollution, thus enhancing accountability and societal pressure, and indirectly resulting in better societal outcomes.

Not every data domain has access to information of the same economic potential. An OECD report from 2006[21] ranked the different data domains according to their commercialization potential. The top of the list features data domains such as Geographic information, Meteorological and Environmental information and Economic and Business Information, while the bottom of the list is composed of Cultural content, Political Content and Educational Content.[22] Similar findings from a range of studies on the value of Open Government Data and on the value of AI in different sectors have been included to distinguish between sectors of high, medium

17. Jetzek, T. et al. 2012: The Value of Open Government Data: A Strategic Analysis Framework
18. Creating value through open data (Carrera, Chan, Fischer, & van Steenbergen, 2015)
19. Information Management and Big Data - A Reference Architecture (Oracle White Paper, 2013; https://www.oracle.com/technetwork/topics/entarch/articles/info-mgmt-big-data-ref-arch-1902853.pdf)
20. The Value of Open Government Data: A Strategic Analysis Framework (Jetzek, T. et al. 2012)
21. Creating value through open data (Carrera, Chan, Fischer, & van Steenbergen, 2015)
22. Creating value through open data (Carrera, Chan, Fischer, & van Steenbergen, 2015)

and low business value for AI solutions on government owned datasets[23] [24] [25] [26]. Moreover, in order to build a business model on government owned datasets, companies need a certain degree of stability and longevity in data collection. Datasets that have been collected over long periods of time and with a strong expectancy of continued collection in the same way have higher value for businesses. Since the primary goal of this project is to create value for businesses – and thus society – the dimension measuring value for businesses have been given extra weight when comparing the assessed datasets against each other.

Finally, for datasets to have cross-Nordic value, language barriers and interoperability aspects need to be addressed so that information resources from different organizations and countries can be combined. The availability of information in machine-readable formats as well as a thin layer of commonly agreed metadata could facilitate data cross referencing and interoperability, thereby enhancing value for re-use considerably. Moreover, in a national context some datasets will be too small to train efficient AI algorithms on. Volume is important where datasets are relatively generic and thus exist and display the same characteristics across the Nordic countries and linking them is the key for developing AI solutions with high business value.

Identifying high-value Nordic datasets for AI solutions

The aim of this project is not to uncover the full universe of data in the Nordic countries, but to find a subset that demonstrates high potential value for AI solutions if made publicly available. The identified subset of datasets is generated based on a combination of desk research and interviews with experts and public data owners in the Nordic countries, including input from the working group on AI and Access to Data in the Nordic Council of Ministers.

To further filter down the initial subset of datasets, a range of selection criteria have been used to exclude datasets from this study. These include an assessment of whether the specific type of data or datasets exist in all or most of the Nordic countries, removing datasets that were unique to one or two Nordic countries due to unique national conditions, and a preliminary assessment of the potential business value of a datasets, especially pertaining to an assumed lack of business demand for data should it be made publicly available.

The final distribution of assessed datasets reflects a political focus on climate and sustainability, a business demand for health data and a desire to provide an initial list of high-value Nordic datasets, including examples of AI relevant datasets across different data categories, that can inspire dataowners and policymakers across the Nordic countries to make government owned datasets publicly available.

23. Analyse af efterspørgsel og markedstendenser inden for offentlige data (Deloitte, 2017; https://data.virk.dk/sites/default/files/analyse_af_efterspoergsel_og_markedstendenser_inden_for_offentlige_data.pdf)
24. Open Growth – Stimulating demand for open data in the UK (Deloitte, 2012; https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/deloitte-analytics/open-growth.pdf)
25. How AI Boosts Industry Profits and Innovation (Accenture, 2017; https://www.accenture.com/fr-fr/_acnmedia/36dc7f76eab444cab6a7f44017cc3997.pdf)
26. Artificial Intelligence in Swedish Business and Society (Vinnova, 2018)

Nordic high-value dataset

This chapter will present the final list of high-value datasets identified in the project, following the method described in the previous chapter.

This project has looked at different *types and use cases of data*, ranging from data domains to specific datasets and specific solutions and models developed on datasets. All datasets have been assessed based on a representative dataset from one of the five Nordic countries. For example, the assessment of Weather Data is based on information gleamed from the Swedish Weather Data. For the results to be valid across the Nordic countries, it is assumed that similar datasets exhibit similar characteristics across the Nordic countries.

It is also important to note that data does not need to be publicly available in all Nordic countries in order to be classified as a Nordic high-value dataset. Datasets in one Nordic country can be of high value for re-users in another Nordic country, and some Nordic datasets are large and rich enough in themselves to be valuable without being linked to or augmented with other Nordic datasets.

Table 1 on the next page presents the initial ranked list of Nordic high-value datasets. The top-ranking datasets, Groundwater, Weather Data, Road Cameras (Photos), Traffic Events and Roadworks, Energy Data (Aggregates) and Company Announcements, have been classified as having a Very High value for developing AI solutions in the Nordic countries.

A more detailed description of each of the assessed datasets can be found in Appendix 2, including short descriptions of barriers and actions related to the dataset in question.

Estimated value for businesses

Due to the way datasets have been selected, most of the assessed datasets have a high or very high **estimated value for businesses**. These are datasets that have been collected in the same way for a long time by public organizations and where it is expected that the datasets are being collected and published in the same way for a long time going forward.

Moreover, these are also dataset within data domains and/or sectors of the economy where open data and AI have been proven to or are strongly expected to generate high value, such as Geospatial, Environment, Mobility and Health.

The solutions on the list, the Danish Nature Recognition dataset and the Building Data (Photos), have no history of prior collection of data and do not appeal to companies wanting to build a business model on this basis.

The datasets on Spoken Language and Written Language have only been assigned a *Moderate* score on the value for business dimension. This is because datasets within Culture and Arts historically are used less for AI development than datasets from

other data domains. However, recent advancements within e.g. natural language processing could make datasets from this data domain highly relevant for businesses in the coming years.

Barriers for value-realization

The top-ranking datasets only have low or moderate **barriers for value-realization**. These datasets are characterized by having no obvious legal barriers that could block the process towards making the dataset available to the public, e.g. by not containing sensitive information.

Generally, these datasets are also characterized by low assumed costs associated with making the datasets publicly available and maintaining them once they are public[27].

For the Road Camera datasets, the data is continuously collected from roadside cameras. This data is viewable in all the Nordic countries and open and accessible in three, implying that barriers for making the Road Camera datasets public are surmountable in the Nordic countries, where the datasets still lack to be made publicly available. Similarly, for the dataset on Traffic Events and Roadworks, public organizations in some of the Nordic countries are already making this data available in formats conducive for AI.

For the majority of the datasets, it is clear who is responsible for the dataset. Unclear or mixed responsibility of data ownership can be a barrier for openness. This is partly true for datasets being constructed as part of research projects, e.g. the datasets on Spoken Language and Written Language, and datasets being collected and published by another public organization than the one that constructed them, e.g. the Biobank register data.

27. In the assessment, costs refer solely to the work related to preparing the dataset for publication and not the costs associated with technical infrastructure, storage costs, etc.

Table 1: Summary of assessed datasets

Dataset	Example country	AI-relevance	Barriers*	Societal value	Estimated value for businesses	Cross-Nordic value	Summary score
Groundwater	SE	Very High	Low	Moderate	Very High	Very High	Very High
Weather data	SE	Very High	Moderate	Very High	Very High	Very High	Very High
Road cameras (photos)	FI	Very High	Low	High	Very High	High	Very High
Traffic events and roadworks	IS	High	Low	High	Very High	High	Very High
Energy data (aggregates)	DK	High	Moderate	High	Very High	High	Very High
Company announcements	NO	High	Low	Moderate	Very High	High	Very High
Flooding	FI	High	Low	Moderate	Very High	Moderate	High
Work accidents	FI	Very High	Moderate	Moderate	Very High	Moderate	High
Company specific data	SE	High	Moderate	Moderate	Very High	High	High
Energy data (individual level)	DK	High	Moderate	High	Very High	Moderate	High
Air quality	FI	Very High	Low	Moderate	High	High	High
Cancer registry	IS	High	High	Moderate	Very High	High	High
Biobank register	DK	High	High	Moderate	Very High	High	High
Rheumatological data	DK	High	Moderate	Moderate	Very High	Moderate	High
Area management	NO	High	Low	Moderate	Very High	Moderate	High
BioImages	SE	High	Moderate	Moderate	Very High	Moderate	High
Bankruptcy	FI	Moderate	Moderate	Moderate	Very High	Moderate	High
Surface water	NO	High	Moderate	Moderate	Very High	Moderate	High
Regulation plans	NO	High	Low	Moderate	Very High	Moderate	High
Written language	IS	High	Low	Moderate	Moderate	High	High
Data on product tests	DK	High	Low	Moderate	High	Moderate	High
Building data (photos)	DK	High	Moderate	Moderate	Very High	Moderate	High
Waste	DK	High	High	Moderate	Very High	Moderate	Moderate
Spoken language	IS	Moderate	Low	Moderate	Moderate	Moderate	Moderate
Nature recognition	NO	Moderate	Low	Moderate	High	Moderate	Moderate

Notes: An explanation of the different dimensions can be found in Appendix 1. Further details on the datasets and their scores can be found in Appendix 2. *Note that Barriers are scored differently from the other dimensions, e.g. making Low Barriers equivalent to Very High in the other dimensions.

Cross-Nordic value

Only two of the datasets on the list have been assigned a score of Very High **cross-Nordic value**. This is the Groundwater data and the Weather data. Both datasets are very similar across the Nordic countries with respect to variables and information, which would enable re-users to merge datasets from different countries without too much effort. For the Weather data, the cross-border nature of that type of data also contributes strongly to a high cross-Nordic value. These datasets are also characterized by a high degree of openness of data across the Nordic countries, enabling possibilities of strong Nordic collaboration.

Datasets like Company specific data, Air quality, the Biobank Register and the Cancer Registry also score High on cross-Nordic value. These are all datasets with assumed high added value when linked with similar datasets across the Nordic countries. The health register datasets, having access to larger, combined Nordic register datasets, makes it possible for researchers and companies to identify unique correlations and develop unique solutions that would not have been possible based on purely national datasets. It is no coincidence that Nordic health registers is the subject of previous and ongoing Nordic collaboration efforts[28] [29] [30] [31].

Especially for aggregated datasets there is much to gain from Nordic collaboration and accessibility of datasets in all the Nordic countries. A good example is the dataset on Work Accidents. Because the dataset in its raw form contains sensitive information on individuals, their occupation and sickness history, it is aggregated before it is made publicly available, reducing its value and relevance for AI solutions significantly. However, having access to datasets on work accidents for all the Nordic countries would increase the volume of the aggregated data by a factor five, making data much more relevant to apply AI algorithms and applications on.

Datasets with a *Moderate* score on cross-Nordic value still relevant for Nordic collaboration efforts. The way the criteria on cross-Nordic have been developed, datasets receive a high score on cross-Nordic value if Nordic collaboration and merging of similar datasets across the Nordic countries is deemed to be a necessary requirement for creating value through the development of AI solutions, or if the datasets contain information that naturally reaches across borders, which is the case for e.g. weather data, air quality data and some types of mobility data. Cross-Nordic value is thus an indicator of which datasets have the highest potential value associated with joint Nordic actions.

Societal value

Besides generating value for businesses, making datasets publicly available can positively benefit society by helping achieve a range of societal goals. The datasets with a high or very high **societal value** are datasets that are considered useful in AI

28. A vision of a Nordic secure digital infrastructure for health data: The Nordic Commons (NordForsk, 2019)
29. NOS-M Report: Personalised Medicine in the Nordic Countries (NordForsk, 2019)
30. Joint Nordic Registers and Biobanks - A goldmine for health and welfare research (NordForsk, 2014)
31. Nordic Innovation program on Health, Demography and Quality of Life (https://www.nordicinnovation.org/health)

applications or AI solutions for a greater good, such as having a positive impact on climate change and improving the health and well-being of Nordic citizens.

Weather data is vital for both businesses and society. Weather forecasts are important for e.g. agriculture and road transport, and rain forecasts play an important role in e.g. wastewater treatment. Historically, weather data affects city planning, the building industry and more.

The Road Cameras datasets and the Traffic Events and Roadworks dataset could be used by businesses to help limit congestion on the roads and reduce CO_2-emissions. These datasets could also help prevent or reduce the number of traffic accidents and make it safer on the roads. Similarly, the Energy datasets could be used to improve energy efficiency and promote green energy consumption.

The assessed Health datasets have only been assigned a moderate score on societal value. This is due to the low number of different societal goals they can be used to achieve. However, all of them are expected to be of great importance for society with respect to innovative treatment of diseases, empowerment of patients through increased information about sickness and symptoms and improving efficiency in the health sector.

Even the lower-ranking datasets on the list, such as Nature Recognition, Spoken Language and Waste, are each expected to be able to achieve societal goals. The Nature Recognition datasets can be used to monitor and protect biodiversity; the Spoken Language dataset can be used for educational purposes; and the dataset on Waste can be used for solutions within circularity and the green transition.

AI-relevance

Finally, the datasets with a high or very high **AI-relevance** are the ones having the specific characteristics needed to be used in the development of AI solutions, such as size, richness and potentially labels or similar.

The majority of the assessed datasets are tabular data with structured values in rows and columns. However, the datasets on Spoken Language, Written Language, Road Cameras and BioImages contain non-tabular data in the form of audio clips, text, and images. The AI technology is uniquely adapted to handle these types of unstructured data formats. Making more of these types of data available to the public would be of great value for companies wanting to develop innovative new solutions. Very often, however, these types of data are not considered very valuable to the organizations that own them, since they do not know what to do with them. As organizations mature and AI competencies become more common in the public sector, one would expect that unstructured datasets are more likely collected, used and subsequently made publicly available.

Another important dataset feature for AI purposes is the presence of labels or ground truths in the dataset. For the Nature Recognition dataset, labels indicate which type of nature a given photo illustrates. For the BioImages dataset, labels provide information on what can be interpreted from the image. Similarly, for the dataset on Spoken Language, text accompanying the audio clips connects audio with meaning and enables technologies such as speech-to-text or text-to-speech.

Finally, some of the highest scoring datasets are datasets collected by sensors or similar machine-operated devices. This is the case for the Groundwater data, Weather data, Energy data, Air quality data and the Road Cameras data. Sensors

typically generate vast amounts of data with a high update frequency and are not prone to human-induced dataset errors and interpretations, all of which is of high value for AI solutions. As the public organizations in the Nordic countries become more digitalized, more of these sensor datasets will be collected. Thus, it is important for Nordic policy makers to be aware of the high value associated with these types of data.

Cross-Nordic openness of assessed datasets

Besides identifying and assessing the Nordic high-value datasets presented in the previous section, part of the project has also been conducting a cross-Nordic openness assessment of the identified datasets.

In the context of the project, datasets have been classified as either *Open*, *Difficult to access* and *Closed*. An *Open* dataset is easily found and one can either download the dataset or access it through an API or similar at no or only marginal cost. A *Difficult to access* dataset is characterized by being either difficult to find and/or only accessible at some cost. This is a grey zone category where datasets to a large degree are easy to find but not accessible or re-useable for a variety of reasons, e.g. that companies need to pay to access the data or that only researchers and research institutions can access the dataset for free. It can also be datasets that are presented as open datasets but where there are no easy ways for companies to get direct access to the dataset, e.g. that datasets are shown on a website but there are no download options. Finally, a *Closed* dataset cannot be found or is not accessible.

Overall, the figure shows that there are many opportunities for the Nordic countries to increase the degree of openness in high-value data domains.

Figure 1 on the following page shows the openness status of the assessed datasets across the Nordic countries.

Most datasets can be found as open datasets in at least one and often several of the Nordic countries. Thus, ample opportunities are present in all data domains for strong Nordic collaboration. The Nordic countries are already cooperating on sharing research data and the work involved in facilitating a Nordic research e-infrastructure.[32] Data sharing projects have also been conducted on health datasets, the latest resulting in a report from Nordforsk on how health data from individual Nordic countries securely can be shared and/or combined across borders[33]. This report concludes that the data sources in the Nordic countries constitute a unique gold mine not available anywhere in the world but that there is a risk that this resource will be lost unless made more easily accessible.

In the data domains of Geospatial data and Environmental data, there is a very high degree of openness across the Nordic countries. This is partly due to EU legislation (e.g. the INSPIRE Directive) and partly due to these datasets being classified as *Basic Data*, essential national datasets containing high quality information[34] [35].

Health datasets are typically closed and not accessible across the Nordic countries, but there are notable differences. The Work Accidents dataset (and datasets with

32. The State of Open Science in the Nordic Countries: Enabling Data Science in the Nordic Region (NordForsk, 2018)
33. A vision of a Nordic secure digital infrastructure for health data: The Nordic Commons (NordForsk, 2019)
34. Good Basic Data for Everyone – a Driver for Growth and Efficiency (The Danish Government / Local Government Denmark, 2012; https://en.digst.dk/media/14139/grunddata_uk_web_05102012_publication.pdf)
35. Uppdrag om saker och effektiv tillgang till grunddata (Finansdepartementet, 2018)

similar content) have been made easily accessible in Denmark, Norway and Sweden.

Overall, the figure shows that there are many opportunities for the Nordic countries to increase the degree of openness in high-value data domains.

Figure 1 – Openness of assessed datasets across the Nordic countries

The next two sections goes into further detail on two of the highest ranking datasets, Groundwater and Road camera data (photos), and what needs to be done in each of the Nordic countries in terms of making the datasets publicly available and/or usable for AI in order to realize their potential value.

Case: Groundwater data in the Nordic countries

Groundwater data has been selected as a case because of its high value and the low barriers associated with making this type of data more accessible for companies across the Nordic countries.

Groundwater data consists of a range of datasets on specific information related to national groundwater aquifers. Monitoring of e.g. groundwater quality has been undertaken in most European countries since the 1970s and 1980s[36], but there are large differences between countries as to what information is gathered and what is made publicly available. Moreover, most of the Groundwater datasets rely on samples gathered at the different aquifers and then analyzed in a laboratory. The implication is that updating these datasets is a costly and time-consuming process, and there are large differences in the sampling frequency in the Nordic countries. The following are examples of exisiting Groundwater datasets in the Nordic countries:

- Observation points: Bored and dug wells
- Groundwater level, temperature, spring level and spring discharge
- Chemical groundwater analyses
- Hardness of drinking water
- Quality of drinking water

The Groundwater datasets are collectively assessed to be of high value for AI development. The quantity and quality of the data is significant, with many million entries and many variables (location, depth, minerals, chemical constituents etc.). Many of the datasets have been available since the 1970's and have been digitized before computer records. For businesses, the datasets have the history and longevity necessary for building stable AI applications. The monitoring of groundwater resources is covered by EU legislation[37], ensuring a certain degree of sameness and comparability of groundwater datasets across the Nordic countries.

Publishing the datasets does not prejudice GDPR legislation and the extra costs associated with making the datasets publicly available and maintaining them are small, since the responsible public organizations continuously are producing and working with the datasets irrespective of data being published. For this case, focus has been on Quality of drinking water datasets.

The table below summarizes the key parameters relevant to Quality of drinking water datasets across the Nordic countries. The table contains information on the responsible organization, the openness of the dataset, whether the dataset is accessible through an API, whether metadata has been published in a machine-readable format and finally the language the dataset is available in.

36. Groundwater monitoring in Europe, https://www.eea.europa.eu/publications/92-9167-032-4
37. Groundwater monitoring in Europe, https://www.eea.europa.eu/publications/92-9167-032-4

Table 2: Quality of drinking water datasets across the Nordic countries

	Responsible organization	Openness	API accessible	Metadata	Language
Denmark	The Geological Survey of Denmark and Greenland (GEUS)	Available through the Jupiter database[38]	-	Not available	Dataset not available. Other groundwater datasets published in Danish and English
Finland	Finnish Environment Institute (SYKE)	Available through SYKE's Open Data platform[39]	API accessible	Available (xml)	Available in Finnish, Swedish and English
Iceland	Icelandic Food and Veterinary Authority (IFVA)	Not published	-	Not available	Dataset not available
Norway	The Geological Survey of Norway (NGU)	Available through the ngu.no data service[40]	API accessible	Available (xml)	Available in Norwegian and English
Sweden	Geological Survey of Sweden (SGU)	Available through the sgu.se data service[41]	API accessible	Available (xml)	Available in Swedish and English

Notes: [38] [39] [40] [41]

Concluding from the table above, the following next steps for making Quality of drinking water datasets more accessible for companies and AI development across the Nordic countries are:

Denmark:

- Data should be made available through an API.
- Metadata should be published alongside the dataset and mirror the metadata published in Finland, Norway and Sweden.

Iceland:

- Data on drinking water should be made publicly available through an API.
- Metadata should be published alongside the dataset and mirror the metadata published in Finland, Norway and Sweden.

The Finnish, Norwegian and Swedish datasets on drinking water quality could be made more visible for companies, e.g. by linking to datasets on National Open Data platforms. Moreover, work could progress with publishing the metadata in English, so it is easier to understand and access for companies in the other Nordic countries.

38. https://www.geus.dk/produkter-ydelser-og-faciliteter/data-og-kort/national-boringsdatabase-jupiter/adgang-til-data/data-gennem-pcjupiter-og-pcjupiterxl-format/
39. https://www.syke.fi/en-US/Open_information
40. https://www.ngu.no/grunnvanninorge/
41. http://resource.sgu.se/service/wms/130/miljoovervakning_grundvatten

Addressing the abovementioned recommendations at a national level would benefit companies seeking to develop AI applications on datasets on drinking water quality in all the Nordic countries. For these companies, having access to similar datasets across the Nordic countries would allow them to build stronger solutions on more data and for a larger potential target group.

Case: Road camera data (photos) in the Nordic countries

Road camera data has been selected as case because of its high value and the low barriers associated with making this type of data more accessible for companies across the Nordic countries.

Road camera data consists of a photo stream or data feed from webcameras located alongside roads in the Nordic countries. The cameras provide information on current traffic flow and weather conditions.

The road camera datasets are assessed to be of high value for AI development. The quantity of photos taken is significant, and data feeds are close to being real-time. Moreover, data across the Nordic countries is similar and clear data formats facilitate linking and using data from different countries. Most of the data is available in DATEX II (specification for DATa EXchange between traffic and travel information centres) format, which is an European standard for exchange of traffic information[42].

Mobility is an area with large business demands for data and characterised as a well-developed market for applications and solutions. Road camera data could be used by freight or delivery services to follow traffic conditions and plan routes accordingly. This could reduce congestion and CO2-emissions. Radio stations can add a live camera feed to a traffic news page, and organizations with staff intranets could add the traffic camera feed so people can plan their journey before leaving. Data exists in all the Nordic countries and has high longevity.

The table below summarizes the key parameters relevant to Road camera data across the Nordic countries. The table contains information on the responsible organization, the openness of the dataset, whether the dataset is accessible through an API, whether metadata has been published in a machine-readable format and finally the language the dataset is available in.

All countries have agencies that capture and use road camera data, as should be evident from the table below.

42. https://www.datex2.eu/

Table 3: Road camera data across the Nordic countries

	Responsible organization	Openness	API accessible	Metadata	Language
Denmark	Vejdirektoratet[43]	Data can be accessed by contacting the agency and paying a small bi-annual fee	Data feed	Not available	Danish
Finland	Digitraffic[44]	Data can be accessed through the Digitraffic API	API	XML/JSON	English
Iceland	Vegagerðin[45] (Icelandic Road and Coastal Administration)	Data can be accessed through an API at the Vegagerðin website	API	XML	Icelandic
Norway	Statens Vegvesen[46]	Data can be accessed through the dataportal at Statens Vegvesen. Access requires login	API	Unclear, requires login	Norwegian
Sweden	Trafikverket[47]	Data can be accessed through the Trafikverket API. Access requires login	API	XML, JSON	Variables in English, the rest in Swedish

Notes [43] [44] [45] [46] [47]

43. https://www.vejdirektoratet.dk/side/viden-om-ydelser-trafikinformation-som-data
44. https://www.digitraffic.fi/en/road-traffic/#weather-camera-image-history-for-the-last-24-hours
45. http://gagnaveita.vegagerdin.is/api/vefmyndavelar2014_1
46. https://dataut.vegvesen.no/dataset/webkamera
47. https://api.trafikinfo.trafikverket.se/API/Model

The implementation of GDPR in the different Nordic countries have different implications for the road camera datasets. In Denmark, old photos should continuously be replaced with updated versions (every 5th second) and re-users are not allowed to save and use old photos. In Sweden, old photos are also replaced as soon an updated camera image comes in. Conversely, in Finland, the photo history is available for up to 24 hours. Another issue is being able to identify individuals. In Norway, re-users are obliged to contact the agency if individuals or registration plates can be seen from the photos.

Concluding from the above, the following next steps for making Road camera datasets more accessible for companies and AI development across the Nordic countries are:

Denmark:

- Data could be made freely available through an API. Work is currently underway in this area and could be supported.
- Metadata should be published alongside the dataset and mirror the metadata published in Finland, Norway and Sweden.

Iceland:

- Data is only available on the Icelandic version of the website and is therefore difficult to find for non-Icelandic re-users.

In general, the Nordic countries are good at making road camera data available for presentation on their websites but options for download and access to data could be made clearer and should optimally be available on the same webpage.

Moreover, different interpretations of GDPR regulation with respect to these types of data could prove an issue for companies wanting to use road camera data from different Nordic countries. To avoid this, work could go into harmonizing the interpretations and provide a common Nordic framework for making road camera data available, including funding to develop software and/or algorithms that can blur out individuals and registration plates, thus preventing GDPR concerns and -issues.

Addressing the abovementioned recommendations at a national level would benefit companies seeking to develop AI applications on datasets with road camera photos in all the Nordic countries. For AI companies, having access to similar datasets across the Nordic countries would allow them to build stronger solutions on more data and for a larger potential target group. The added variety in road and weather conditions that comes from collecting and linking road camera datasets from across the Nordic countries is also of great value for the companies by making the information in the datasets better suited for AI algorithms.

Barriers and recommendations

This chapter describes the identified **barriers and recommendations for AI-utilization of datasets** across the Nordic countries.

There are several opportunities for improvement in order to make more data publicly available for AI solutions in the Nordic countries, both short-term and long-term. Some of these are better handled and addressed at national level, while several are ideal for joint Nordic action and collaboration. The suggested joint Nordic actions and opportunities for collaboration are presented in chapter 4.

This report identifies two prime opportunities:

1. Government owned datasets can be made more visible to companies, generating interest in datasets and demand for making datasets publicly available. This can help the public sector identify which datasets to make public first. Visibility can be furthered through hackathons, promoting the Nordic open data portals and encouraging public organizations to publish information on datasets that have yet to be made publicly available.

2. Government owned datasets can be made more available for AI solutions and development (AI readiness) by providing easier access through e.g. APIs, releasing metadata and dataset descriptions alongside the datasets, and if possible, ensuring dataset interoperability between the Nordic countries.

The opportunities above cover several of the recommendations described on the following pages.

The recommendations are grouped according to whom they are relevant for.

Recommendations targeted...

The Nordic/National level
Public organizations, working groups and policy makers that are involved at the strategic development of Nordic data collaboration, either through Nordic collaboration or through national initiatives.

The data generators
Public organizations that own, collect and generate datasets.

The data publishers
Public organizations that make their own or datasets from other public organizations available to the public.

The data re-users
Private companies and citizens that use government owned datasets to create new solutions or applications. Could also refer to the public sector when re-using data from other public organizations.

Owerview of barriers and recommendations

The table below provides an overview of the identified barriers/challenges, how to address them and to whom the recommendation is addressed. Following the table, the recommendations for each target group is further described in detail in separate subsections. The Nordic level is addressed in the following chapter.

Barriers	Recommendations	Target group
A: Making data publicly available is often not prioritized enough	A1: Collect or construct commendable showcases and examples of the value of government owned open data from across the Nordic countries	National level
B: Publicly available government datasets might not be re-useable for AI solutions	B1: Use data format recommendations and standards; in particular international standards when available B2: Facilitate emergence of data ecosystems	National level
C: Lack of a volume-based market with sizeable business value	C1: Exemplify needs and avoid building proprietary solutions, whenever possible C2: Find ways of funding to compensate public organizations that are publishing data at a cost for businesses C3: Enlist the help of citizens, startups and the open data community C4: Encourage and support collaboration with startups and SMEs	National level
D: Datasets contain sensitive information on individuals	D1: Fund projects investigating fully GDPR compliant options for releasing sensitive information D2: Encourage public organizations to release sensitive datasets in an aggregated form	National level
E: Lack of overview of internal data resources	E1: Ensure that data management and the data architecture promote easy overview of and access to data	Data generators
F: The quality of datasets in the organization are often perceived as not being high enough	F1: Enable data re-users to help improve dataset quality F2: Make datasets available with documentation of the processes that were used to create a specific dataset	Data generators

G: Publishing data can be time-consuming and costly	G1: Facilitate making data publicly available through a standardized data submission setup G2: Encourage data generators to utilize professional data publishers	Data publishers
H: Companies have limited knowledge about which datasets are collected, created and/or published by public organizations in the Nordic countries.	H1: Create visibility for the datasets that can be made publicly available H2: Promote the open data portals in the Nordic countries H3: Undertake preliminary work into the creation of a cross-Nordic open data portal	Data re-users
I: Companies might not be aware of which solutions there is a public sector demand for	11: Communicate the purpose for which the datasets have been collected I2: Engage in public-private dialogues with the market	Data re-users
J: Datasets might lack metadata and dataset descriptions	J1: Publish datasets with proper and detailed data descriptions J2: Provide guidance for data publishers on the European metadata specification DCAT-AP	Data re-users

National level

This section focuses on recommendations targeted the national level. This entails work on many levels, from supplying the right infrastructure to creating engagement for open data among citizens, society, businesses and in the public sector itself.

Barrier A: Making data publicly available is often not prioritized enough

There is often a lack of knowledge about the value of open data in public organizations, and especially with regard to the potential value generation of AI solutions on government owned datasets. This results in a lack of funding and not enough prioritization of resources and time in governmental agencies.

Recommendation

1. Collect or construct commendable showcases and examples of the value of government owned open data from across the Nordic countries. It is especially important to highlight the value of data for use outside the initial purpose of collecting it (data re-use). Focus should be on exemplifying the potential societal gains associated with the dataset in order to link the open data agenda to the core purpose of the organization.
   a. Showcases can be collected from international studies and/or from governments with strong open data and AI agendas, such as the UK or the US.
   b. Showcases can also be found in the open data community and in civictech applications.

Relevant for the following datasets
Relevant for all the datasets assessed in this project. Less relevant for e.g. weather data, geospatial information and business register data where multiple case studies already have shown the high potential value of making data publicly available.

Barrier B: Publicly available government datasets might not be re-useable for AI solutions

For many government owned datasets, there is a lack of standardized data formats or data access interfaces. Moreover, many datasets that are published lack a cross-national or international perspective. Variables in the dataset, metadata descriptions and similar are often only available in the national language of the data publisher.

Recommendations

1. Use data format recommendations and standards; in particular international standards when available; use CSV file format as a baseline. Follow open data recommendations and standards also for licensing. Keep in line with EU guidelines and practices developed at the European Data Portal[48].
2. Facilitate emergence of data ecosystems. An example of such a data ecosystem is Trafiklab in Sweden, which utilizes public timetable information to add value to travelers, for example connections, cycling routes and safe ways home. Trafiklab is a community for open traffic data. It is a startup-like environment with open data releases being published and hackathons being organized for hands-on experience. A 11-member steering board, comprised of local transport directors, set the direction of the work to ensure it is usefulfor commuters. One internal benefit of building a data ecosystem is that it contributes to an understanding of the usage of data as well as provides a data tradition and/or culture.

Relevant for the following datasets

Data format standards are relevant for all datasets, but less so for datasets within data domains strongly regulated by the EU, e.g. by the INSPIRE Directive, or where clear standards already exist and are extensively used, e.g. for geospatial data and weather data. Facilitating data ecosystems is important in all data domains, and there are good examples from e.g. Sweden on the emergence of data ecosystems in the data domains of mobility, health, public governance and culture[49].

48. A short introduction to open data formats can be found here: https://www.europeandataportal.eu/elearning/en/module9/#/id/co-01
49. Respectively, https://www.trafiklab.se/, https://liu.se/en/research/aida, https://www.vinnova.se/en/p/smarter-city-labs/ and https://www.ai.se/en/projects-7/swedish-language-data-lab

Barrier C: Lack of a volume-based market with sizeable business value

To create AI solutions on government owned datasets, businesses require a volume-based market with a sizeable business value. Public organizations often have a difficult time formulating issues that private companies could solve for them and might not always be keen to do so.

Similarly, some public agencies run data provision services as a business to finance activities and have an economic incentive not to make data publicly available, unless compensated by the national government.

Recommendations

1. Exemplify needs and avoid building proprietary solutions, whenever possible. When making a government-developed solution available to the public, ensure to also publish the raw datasets used to develop the solution in the first place. This way, the government-developed solutions act as inspiration for and not saturation of market possibilities of the datasets. The more the raw data has been aggregated, filtered or analyzed before being made public, the less new things or correlations might be discovered from those data. The Swedish project *JobTech Development* is a good example of publishing raw datasets on employment and job adverts in Sweden alongside inspirations for dataset re-use[50].

2. Find ways of funding to compensate public organizations that are publishing data at a cost for businesses. As shown in multiple business cases, the business and societal value of data being open and free of charge quickly surpasses the initial loss of revenue.
   a. Help public organizations construct business cases to further open data agenda.
   b. Advance the ongoing development of national infrastructure and recommendations for open data; encourage its use.

3. Enlist the help of citizens, startups and the open data community. Hackathons create visibility of datasets and illustrate their value and potential for re-use, also spurring a business demand. The Swedish site Challengesgov.se is one example, a platform developed to promote open and data-driven innovation by publishing current societal challenges and links to relevant open datasets. Public organizations are invited to publish their challenges, typically including details on what users need to target and which kinds of open data that is available. Starting in 2018 as part of a commission from the Swedish government to promote open and data-driven innovation, the platform has so far hosted 17 challenges. The latest and current challenge concerns package services in sparsely populated rural areas, looking for data-driven, user-adapted, scalable and sustainable solutions for the entire supply chain[51].

50. https://www.jobtechdev.se/
51. https://challengesgov.se/sveriges-paketombud-data-challenge/

4. Encourage and support collaboration with startups and SMEs. A good way to invite collaboration with startups and SMEs is for the public authority producing data to identify what challenges they need solved and then invite companies to solve them. The challenges may be published on their own website; also look for hackathons or challenge driven initiatives organized by others (e.g. the open source community) to get better coverage. The current COVID-19 situation provides good examples of a challenge driven innovation to participate in, see e.g. the initiative *Tackling coronavirus (COVID-19)* started by OECD[52]. Participation in match-making events for startups is also a possibility, e.g. through platforms established for that purpose, e.g. Ignite Sweden[53]. An obstacle both for data providers and small companies is how to find funding but also *more muscle*, i.e. relevant partners. Here, governmental funding agencies may endorse collaboration through directed support. Swedish Vinnova provides a good example with many collaboration programmes, notably the Datalab programme[54]intended to gather many actors and creating domain specific platforms making data public and ready to be used e.g. for AI.

Relevant for the following datasets

Providing access to raw datasets is especially relevant for the two solutions in this project; the Danish Nature Recognition dataset and the Building Data (Photos) dataset. Funding opportunities and hackathons are relevant for all the datasets, but especially for datasets within the data domains of health, culture and public governance, where there is less of a tradition for providing data access compared to e.g. geospatial datasets and mobility data.

52. http://www.oecd.org/coronavirus/en/
53. https://ignitesweden.org/public
54. https://www.vinnova.se/en/calls-for-proposals/data-driven-innovation/datalabb-och-datafabrik-som-nationell-resurs-2020/

Barrier D: Datasets contain sensitive information on individuals

One of the major issues preventing datasets from being made available to the public is the risk of disclosing sensitive information related to individuals. This is a barrier for many datasets of high value for businesses and should be addressed by policy makers at the national or Nordic level. The high number of examples of cross-Nordic (research) cooperation on health data registers and the voiced data demand from researchers and companies point to this being a pivotal area to focus on going forward.

Recommendations

1. Fund projects investigating fully GDPR compliant options for releasing sensitive information. These options include, but are not limited to, anonymization, pseudo-anonymization and synthesizing data. There are already projects underway in the Nordic countries, e.g. Synthetic Health and Research Data (SHARED)[55] and Synthetic data from the Norwegian National Register[56]. SHARED is a research collaboration between researchers from Denmark and Finland and the Novo Nordic Foundation. Its aim is to prove that it is possible to transform original health data into synthetic data in a way where it is not possible to identify individuals in the data. Similarly, The Norwegian Tax Agency has provided synthetic register data for integration tests. It is the first step in a cross-ministerial project on synthetic test data in Norway. Further support for these or similar projects could speed up the refinement of these method and make it more accessible for governmental agencies in the Nordic countries.

2. Encourage public organizations to provide access to aggregated datasets. Most datasets can be aggregated to a level where they still create value without conflicting with GDPRand these aggregated datasets still hold high value for businesses. Besides the time and resources spent publishing datasets, aggregating datasets requires knowledge about the potential re-users and their data needs. Good examples of aggregated data re-usage in these fields can be collected across the Nordic countries and be used to inspire further data openness.
   a. Collect and share good examples of highly re-used aggregated datasets.
   b. Explore re-user demand for data to use agency time and resources for maximum impact.

Relevant for the following datasets
Relevant for the Biobank register, BioImages, Cancer Registry, Energy data (individual level), Rheumatological data, Waste, and Work accidents. In general, relevant for all datasets containing sensitive information.

55. https://novonordiskfonden.dk/da/nyheder/syntetiske-sundhedsdata-kan-sikre-bedre-forebyggelse-og-behandling/
56. Et syntetisk Folkeregister - Et samarbeid mellom prosjektet og Testsenteret i Skatteetaten

Data generators

Data generators are responsible for creating, collecting and/or generating the datasets used by public organizations. Even though data generators do not necessarily make their data publicly available themselves, instead relying on other public organizations taking on the role of data publisher, the data generators can facilitate and further the open data agenda through better data management and higher quality data.

Barrier E: Lack of overview of internal data resources

When data resources are not easily accessible even internally, publishing datasets to externals are not prioritized or might not be possible. Many public organizations are still undergoing internal projects regarding data governance, data management and data architecture. Moreover, it is often not feasible in terms of the resources and time that need to go into finding, collecting, quality assuring and publishing datasets.

For data generators, developing the data sharing platform is the least of it, and data management and getting data into the sharing platform is the hard part. Data creators might not even know that their datasets can be made publicly available if there is no centralized agenda driving open data in the organization.

Finally, if the executive management of a public organization does not drive the open data agenda, there is an increased risk for a deficit of internal funding and resources to establish and maintain open datasets.

Recommendations

1. Ensure that data management and the data architecture promote easy overview of and access to data. It would be beneficial for organizations to work towards creating an overview of internal data resources. This will serve a dual purpose of facilitating data usage internally in the organization while at the same time facilitating making data publicly available. Organizations cannot and will not publish data they do not know exist.
   a. Organisations could benefit from thinking publication of datasets as a part of all projects involving data within the organization.

Relevant for the following datasets
Relevant for all datasets assessed in this project. Most government owned datasets are made publicly available on an ad hoc basis and are not tied to internal efficiency or data management projects.

Barrier F: The quality of datasets in the organization are often perceived as not being high enough

Doubts about the quality of datasets is a barrier for both data generators, data publishers and data re-users. For the data generator, low data quality can prove a large barrier to wanting to publish data, based on worries about the wrong re-use of data and the risk of errors in data being exposed. Ensuring high data quality can be a costly and time-consuming activity, often requiring a specialized unit in the organization responsible for a final quality check before data is made publicly available.

Recommendations

1. Enable data re-users to help improve dataset quality. Making datasets available can be used to ensure high quality of datasets by using external resources to point out dataset shortcomings and sometimes solutions. Moreover, since high-value datasets often are datasets collected for a reason related to the purpose of the public organization and thus used often, it is in the public organization's best interest to divert funds into a more rigorous quality checks of internal datasets.
   a. Identify datasets that are essential and core to the public organization. Ensuring high quality of these datasets will prepare them for publishing while also positively benefitting the organization.

2. Make datasets available with documentation of the processes that were used to create a specific dataset. This can also be referred to as *provenance metadata* and is an important part of FAIR data practices[57]. The information helps companies and data generators alike to identify potential bias and areas for workflow and data creation improvements[58].

Relevant for the following datasets
Mostly relevant for datasets where no clear examples of successful publishing are available across the Nordic countries. That is the case for Company generated waste, Data on product tests and the majority of the health datasets. In general, information about data generation processes are lacking for public datasets in the Nordic countries, inhibiting data re-use.

57. The State of Open Science in the Nordic Countries (Anders O. Jaunsen on behalf of NEIC, 2018)
58. A similar point are made in: https://www.nordforsk.org/2018/state-open-science-nordic-countries-enabling-data-science-nordic-region

Data publishers

Data publishers are public organizations supporting data generators in making their data publicly available. Some public organizations will take on the role of both data generator and data publisher, facing both sets of challenges.

Barrier G: Publishing data can be time-consuming and costly

Publishing data for (unknown) re-use outside the data generation organization requires resources in terms of both man-hours and computer systems, resources that will not directly contribute to the everyday work of the organization. Data is most likely not directly ready for publishing but needs to be described, cleaned, checked, quality assured, reformatted and re-structured. New computer and information systems and services must also be set up and managed.

Smaller public organizations often lack a combination of the competencies and resources to create and run an efficient and stable publishing activity. Competencies are lacking in terms of technical skills needed to set up the data infrastructure necessary for making datasets publicly available as well as in terms of the understanding of how to publish and present datasets so that they are highly re-useable for AI solutions.

Moreover, setting aside time and resources for making data available is not likely to happen unless there is either political pressure or strong, voiced data demand from companies, both of which is tied to an understanding of dataset value and how data can create value by being made publicly available.

Recommendations

1. Facilitate making data publicly available through a standardized data submission setup. Optimally, the data submission setup is integrated with the organization's internal information architecture.
   a. Study characteristics of the organizational and information architectural setup within organizations that have long experience with making data publicly available.
   b. Share good examples and guidelines from public organizations in the Nordic countries that other organizations can learn from[59].

2. Encourage data generators to utilize professional data publishers. Not all public organizations should make their own data publicly available. The IT infrastructure and storage can be handled by data publishers with more resources.

59. E.g. in Norway: https://www.difi.no/fagomrader-og-tjenester/digitalisering-og-samordning/digitaliseringsradet/laer-av-andre/difi-deling-av-data

Relevant for the following datasets

This is relevant for all datasets and especially for datasets owned by smaller public organizations that may not have the resources to prioritize making datasets publicly available unless internal efficiency gains or external value generation can be proven. The recommendations are also of high relevance for research datasets such as the BioImages (SCAPIS) dataset; research funding is granted for collecting data and for designing a suitable infrastructure, but not for planning, hosting and funding the long-term management of the dataset.

Data re-users

Data re-users can be both private companies and public organizations using datasets made publicly available for purposes other than what they were collected for. Barriers for re-users need to be addressed by public organizations and Nordic policy makers in order to create value through openness.

Barrier H: Companies have limited knowledge on which datasets are collected, created and/or published by public organizations in the Nordic countries

For companies to either find the specific dataset needed as input to an AI application or getting inspiration for developing new AI solutions and applications, a central and easy way to find overviews of government owned datasets in and across the Nordic countries is needed. For companies, demand for datasets is closely linked to the availability of datasets, implying that business demand by itself is a lackluster indicator for which datasets to prioritize to make publicly available.

Recommendations

1. Create visibility for the datasets that can be made publicly available, potentially spurring business demand that can lead to political focus and increased funding for publishing activities in the organization.
   a. Arrange events, e.g. Hackathons, with datasets that have not yet been made publicly available to create awareness of both which datasets the publishing organization have and what value could be created if those datasets were made publicly available.

2. Promote the open data portals in the Nordic countries. All the Nordic countries have established and are continuously developing their national open data portals. There are ongoing projects on how to automatically communicate publicly available datasets to the EU open data portal. Efforts should be made to more efficiently communicate the existence, use and value of these portals to the business community.
   a. Efforts should go into adding datasets to these portals, creating higher visibility of datasets across the public sector.
   b. Efforts should go into publishing information about datasets that are not publicly available, potentially creating business demand and arguments for making this data publicly available. Inspiration can be drawn from e.g. the Norwegian *Fellesdatakataloget*[60] that includes status about openness and relevant contact information for both open and closed datasets.

60. https://fellesdatakatalog.digdir.no/

3. Undertake preliminary work into the creation of a cross-Nordic open data portal. This would ensure a higher re-use of datasets from across the entirety of the Nordic countries and help create a cross-Nordic market for AI solutions and applications.

Relevant for the following datasets
Relevant for all datasets, but especially those that currently cannot be found through the national open data portals, e.g. the health datasets, the mobility data and data on energy consumption.

Barrier I: Companies might not be aware of which solutions there is a public sector demand for

For companies to expend the resources necessary to locate and explore use and re-use of government owned datasets, they need to be able to identify potential customers for developed AI applications and solutions. The public sector is a potential large customer for many of the AI solutions developed by private companies on government owned datasets, but the public sector does not have the competencies needed to see the AI possibilities of their datasets and does not communicate its need for AI development and AI solutions to the private sector.

Recommendations

1. Communicate the purpose for which the datasets have been collected to help companies identify potential business cases and to create solutions targeted at solving problems for the public sector.
   a. Publish cross-Nordic examples and showcases of AI solutions developed for the public sector on government owned datasets.

2. Engage in public-private dialogues with the market, e.g. by using existing marketplaces and efforts such as GovTech initiatives, Hackathons and other challenges to advertise the need for solutions.

Relevant for all datasets.

Barrier J: Datasets might lack metadata and dataset descriptios

Published datasets often lack the metadata and/or data descriptions necessary for companies to understand the potential of re-use. Moreover, publicly available datasets in the Nordic countries are commonly only published in the national language, making it difficult for companies in the other Nordic countries to find, understand, link and re-use datasets.

If companies cannot gain access to information about data collection, data representativeness, data quality and detailed data content, the publicly available datasets hold little value for businesses.

Recommendations

1. Publish datasets with proper and detailed data descriptions according to common international practices for open datasets. Refer to the EU and national data portals for guidance and good examples. For research data and its metadata, the FAIR principles must be considered[61]

2. Provide guidance for data publishers on the European metadata specification DCAT-AP, which enables standardized retrieval of metadata to data portals.

Relevant for all datasets.

61. The State of Open Science in the Nordic Countries (Anders O. Jaunsen on behalf of NEIC, 2018)

Suggested joint Nordic actions

This chapter contains suggested actions for Nordic collaboration. The actions build on and support the recommendations presented in chapter "Barriers and recommendations" and are designed to ensure strong Nordic cooperation in the efforts to boost the development of AI solutions on government owned datasets.

The suggested joint Nordic actions are presented in order of feasibility and expected impact on data openness and AI usage in the Nordic region for the benefit of society and businesses.

Action 1: Arrange cross-Nordic hackathons on government owned data

To increase public sector awareness of the value of open government data, and to help create and expand a Nordic market for AI on open government data with sizeable business value, it would be beneficial to arrange cross-Nordic hackathons on government owned data. Hackathons would create visibility of datasets for companies and showcase potential use-cases and the value of the datasets for the participating public organizations and companies alike.

Action 2: Collect and showcase examples of the value of government owned data from across the Nordic countries

Another way to create public sector awareness of the potential value of publishing more government data and making it more accessible for companies is to collect and showcase examples from across the Nordic countries.

Examples can be collected and showcased by a working group in the Nordic cooperation and/or funds could be directed towards projects with the aim of gathering good examples and showcases of beneficial re-use of open government data. The Nordic collaboration could support the establishment of a Nordic case-bank with examples and links to the datasets being used.

Action 3: Fund projects creating an overview of which government owned datasets are highly used and demanded by companies across the Nordic countries

This report has taken a first step in identifying high-value datasets across the Nordic countries but there remains to be established a broader and more complete overview of which datasets that currently are in high demand in the different Nordic countries by AI companies and startups.

Further analysis into which datasets are already seeing high re-use across the Nordic countries and especially which datasets are seeing high re-use in some countries and are inaccessible for companies in the other Nordic countries could help public organizations in the Nordic countries to prioritize publishing data that is known to be used for the development of AI applications and solutions.

Follow-up work could go into constructing business cases on these identified datasets, giving public organizations solid economic arguments for directing funds towards making those datasets publicly available. Another similar approach is to establish Nordic public-private networks on data re-use and access to data, where public organizations and AI companies can have a constructive dialogue on government owned datasets and their business potential.

Action 4: Establish Nordic working group on open data standards and formats including best practices when publishing data

A Nordic working group on open data standards and formats could be put together. The purpose of this working group would be to create an overview for public organizations on which open data standards and formats should be used. This work needs to be aligned with international standards and European (EU) guidelines.

Furthermore, such a working group could gather best practices from the Nordic countries on how to publish data in a way that makes data accessible for companies wanting to re-use the data for the development of AI applications and solutions.

Action 5: Fund projects investigating the potential of new or known methods to publish sensitive data in accordance with GDPR

Many of the datasets of the highest value for AI development in the Nordic countries and beyond contain sensitive information on individuals and thus cannot be made accessible in their raw state. The Nordic cooperation could fund projects addressing this issue, e.g. projects that work towards refining the algorithms necessary to create synthetic datasets and/or projects with a similar purpose.

Since the private sector stands to gain a lot from gaining access to these datasets, the Nordic cooperation could also promote and/or fund possible public-private partnerships.

Finally, as work is already underway in the Nordic countries on this issue, there is a need for knowledge sharing and dissemination at the Nordic level, ensuring that cutting edge technologies, solutions and best practices are visible to public organizations across the Nordic countries.

Action 6: Fund projects to make groundwater data and road camera data more accessible for companies across the Nordic countries

In the project, two of the highest ranking datasets - Groundwater and Road camera data (photos) – have been described with respect to what needs to be done in each of the Nordic countries in terms of making the datasets publicly available and/or usable for AI in order to realize their potential value. This suggested action is included to accelerate that process and ensure that data resources are available across the Nordic countries.

Furthermore, it is not enough just to make these data resources more accessible for companies. Projects must also aim to harvest experiences on whether collecting and

linking of data across the Nordic countries facilitate increased AI usage of datasets, or if another approach to furthering the use of open government data for AI solutions and applications is more feasible and/or effective.

Action 7: Promote the open data portals in the Nordic countries

The Nordic countries already have many high-value datasets available for companies. More work could go into promoting the open data portals in the Nordic countries, both internally in the different countries but also at a Nordic level, making it easier for Nordic companies to find and access data from different countries.

The Nordic cooperation should consider linking to open data portals in the Nordic countries on a Nordic website. Funds could be directed towards identifying all open data portals and repositories of open government data across the Nordic countries.

As a continuation of that work, the Nordic cooperation should consider establishing a working group exploring the possibility of having a joint Nordic open data portal, similar to the European Open Data Portal. Nordic datasets are typically very similar with regard to information, variables and quality and would be easier to group and present on a platform separate to the European Open Data Portal.

Action 8: Collect good practice examples from the Nordic countries on good data governance and data management related to publishing datasets

An issue for many public organizations is the lack of good internal data governance and data management practices. It would be beneficial for these organizations, if the Nordic cooperation funded projects identifying good and best practice examples of data governance, data architecture and data management from public organizations in the Nordic countries experienced with creating, collecting, using and publishing data.

List of definitions

Artificial Intelligence
AI is the ability to perform tasks in complex environments without constant guidance by a user, i.e. it is an autonomous process. AI also possesses the ability to improve performance by learning from experience, hence is also an adaptive process [ElementsofAI.com]. Artificial intelligence is often used as an umbrella term for technologies that enable machines to mimic human intelligence, such as computer vision, language processing and machine learning.

Data domain
Following the terminology employed by the European Data Portal, *data domain* refers to a cluster of datasets related to a common topic, e.g. Mobility or Environment.

Dataset
In this report, a dataset is a collection of data, published or curated by a single agent. Data comes in many forms including numbers, words, pixels, imagery, sound and other multi-media, and potentially other types, any of which might be collected into a dataset[62], e.g. a dataset on Finnish air quality or a dataset on Icelandic weather patterns. Dataset is also used to denote a set of data that might be stored in different files or databases.

Metadata
Metadata is documentation that describes data. It can include content such as contact information, details about observations, abbreviations and codes used in the datasets, version information and much more. The Digital Curation Centre provides a catalogue of some common metadata standards[63].

Open Data
Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike. Data must be in an open format, under open licenses, and provided in a form readily processable by a computer.[64]

Public Sector Information Directive
Also known as the "Open Data Directive", the Directive on open data and the re-use of public sector information provides a common legal framework for a European market for government-held data.[65]

Re-use
Re-use means using public sector information for a purpose other than the initial public task it was produced for.

Solution
In this report, solutions refer to use cases of datasets, where government agencies (or companies) have augmented one or more datasets with labels or similar when developing a solution to a specific issue, e.g. classifying types of nature (Nature Recognition) or predicting the weather for the next week (weather forecasts).

62. undefined
63. undefined
64. undefined
65. undefined

Appendix 1: The assessment framework

Purpose of the assessment framework

The purpose of the assessment framework is to assess the potential value of government owned datasets in the Nordic countries for use in artificial intelligence solutions. Value includes value for society, value for businesses and value of Nordic collaboration.

Many of the criteria in the framework can be evaluated by external experts with general knowledge of data but some of the criteria – especially pertaining to barriers to value realisation – requires some degree of interaction with people with some knowledge of the dataset(s) to be evaluated.

Design principles

This section briefly describes the three design principles underlying the assessment framework:ΩNo direct access to data.

- Political relevance.
- Operationality and transparency.

The assessment framework must be useable even though its users have no direct access to data. As a consequence, it consists of a set of questions that can be answered partly via general knowledge of data and the type of data in question and partly through brief interaction with people closer to the dataset of interest.

The framework is designed for use in a political context. This means that criteria are made to weighted – and reweighted – continuously ensuring that that as the political focus changes so can the focus of the framework.

Finally, the criteria and the scores associated with them are presented in a transparent way, rooted in the literature and intended to be used by technical and non-technical experts alike. Another aspect of the operationality of the framework is that the number of criteria is purposely kept small in order to facilitate its use.

Dataset AI-relevance

Data is the foundation for AI and thus the natural starting point for the assessment framework. Any application of AI will only be as good as the quality of data collected. The criteria measuring dataset AI-relevance are of technical nature, intended to assess how useful data is for AI solutions.

The criteria can be grouped into three groups:

- Size
- Structure
- Quality

Table 1 contains the criteria developed to measure AI-relevance.

Table 1 – Criteria for measuring AI dataset relevance

Size	How many unique subjects/items does the dataset contain?
	How many indicators/variables does the dataset contain that are related to its subjects/items?
Structure	Is the dataset generated by humans or machines (e.g. by sensors)?
	Does the dataset contain "authoritative truths" (e.g. labels)?
	Does the dataset contain a time-dimension (e.g. observations over periods of time)?
	Does the dataset contain location information (e.g. GPS coordinates or addresses)?
	Is the dataset structured as values in columns and rows or does it also contain text, pictures, audio, video or similar?
Quality	How many missing observations are expected in the dataset?
	Is the dataset representative of the area of interest (e.g. does it cover a population of subjects or just a subset of a population)?

Barriers for value realisation

There are generally two perspectives on barriers for value realization. First, there are barriers for governments and governmental agencies related to legal issues, costs and technical competencies. For example, in Denmark, an audit report on open government data by the Public Accounts Committee (Rigsrevisionen) in Denmark concluded that one of the main barriers to making more datasets available to the public was technical competencies in the Danish ministries, how to make data available and in which formats[66]

Second, there are barriers for the users of the data in terms of data access and the quality of data provided. For users of data, poor quality of open government data complicates re-use. If government data is made open but its quality is not sufficient, it acts as a main barrier to re-use. The "cleaning up" time and transformation of data can make the re-use inefficient and costly, surpassing capabilities and capacity of the re-user.

The criteria measuring barriers for value realisation have been grouped into the following categories:

- Legal
- Costs
- Re-use

Based on prior experience, the barriers are strongly linked, implying that it might not be necessary to develop detailed criteria for each of these categories.

Table 2 presents the criteria contained in the categories.

66. Rigsrevisionen 2019: Beretning om åbne data. https://www.rigsrevisionen.dk/media/2105061/sr1218.pdf

Table 2 – Criteria for measuring barriers for value realisation

Legal	Is it clear who owns and holds responsibility for dataset quality, correctness and accuracy?
	Does releasing the dataset in its raw form go against GDPR-regulations? (e.g. collected for a different purpose or contains information on identifiable subjects)
Costs	Will making the dataset available to the public result in a loss of revenue? (e.g. is it already available now, but at a cost)
	What is the expected cost (low, medium, high) of making the dataset available to the public, taking into account preparing the dataset for release (incl. ethical considerations about dataset re-use and risk of subject identification), technical competency in the organisation and similar?
	What is the expected cost (low, medium, high) associated with maintaining and updating the dataset?
Re-use	Can the dataset be made available through an API or does it need to be downloaded from a government website?

Lacking the literacy to understand licences and legislation is a common barrier to re-use. Closely intertwined with this is a gap in knowledge and a lack of confidence, which prevents re-users from further exploiting the potential of a dataset. Moreover, the fear of violating intellectual property rights or the privacy of individuals described in the data have gotten worse after the recent implementation of GDPR[67].

Societal value

It is generally acknowledged that assessments of the AI potential for businesses and society are genuinely difficult. Even more difficult is quantifying AI potential in the context of addressing societal challenges linked to environmental and social challenges[68].

In this framework, the logic behind Societal value is that datasets create value if they can be used to achieve societal goals, and the more, the better. Societal value is divided into three groups of societal goals: Economic, Social and Sustainable.

Table 3 – Criteria for measuring Societal value below presents the societal goals chosen in each of the groups.

67. European Data Portal 2018: Analytical report # 11: Re-use of PSI in the public sector.
68. Artificial Intelligence in Swedish Business and Society (Vinnova, 2018)

Table 3 – Criteria for measuring Societal value

Economic	**Employment**
	Government Efficiency
	Growth
Social	Livelihood
	Health
	Education
	Culture
Sustainable	Biodiversity
	Carbon Mitigation
	Air Quality
	Circularity
	Energy Efficiency

In the Open (Government) Data literature, more open data on government operations are believed to improve the quality of a democracy in a country, simply by letting non-governmental agents in a country monitor what the government is doing and how[69]. The approach taken to societal goals in this framework also builds on this assumption, assuming that (more) data on e.g. air quality can help citizens and companies monitor what the government is doing to reduce air pollution, increasing accountability and societal pressure and thus indirectly leading to better outcomes.

In the current political climate in the Nordic countries, some of these societal goals are more politically salient than others, especially in the context of artificial intelligence solutions. As previously described, weighting the different societal goals accordingly can make the framework reflect the political reality, ensuring described datasets are relevant in the current political context. With this in mind, the societal goals **highlighted** in Table 3 have been deemed highly salient and will be weighted (and scored) accordingly.

Estimated value for businesses

The potential value for businesses of a government dataset that have not yet been made publicly available is difficult to assess. Value for businesses is sometimes called the "commercial potential" and has been studied extensively in relation to the broader field of Open Government Data but only with respect to types of data and sectors[70].

Access to new datasets can create value for AI-businesses through several paths:

1. The use of data creates new business models
2. New products are developed using the data

69. The Value of Open Government Data: A Strategic Analysis Framework (Jetzek, T. et al. 2012)
70. Creating value through open data (Carrera, Chan, Fischer, & van Steenbergen, 2015);Open Growth – Stimulating demand for open data in the UK (Deloitte, 2012); How AI Boosts Industry Profits and Innovation (Accenture, 2017); Turning AI into concrete value (Capgemini, 2017); The State of AI 2019 – Chapter 7: Europe's AI startups (MMC Ventures, 2019)

3. Data is used to increase sales and/or attract new customers
4. Data can be used to increase internal business efficiency

In the assessment framework presented in this short report, value for businesses is measured with four criteria. **Table 4** presents these.

Table 4 – Criteria for measuring value for businesses

Commercial value	Does the dataset characterize as a type of data that has low-, medium- or high commercial value?
Indicative market potential	What is the extent of the geographical scope of the dataset?
Availability over time	Looking back: For how long has the dataset been collected (and structured as it is at present)?
	Looking forward: For how long is the dataset expected to be collected (and structured) this way?

These criteria require further explanation. Based on previous reports on the value and use of Open (Government) Data and reports on the expected impact of AI technologies on sectors of the economy, types of data (data categories) can have either low-, medium- or high commercial value.

To illustrate, the following statements hold true for datasets with high commercial value:

1. The dataset can be used in sectors where there is a large market for open data
2. There is a high existing commercial re-use of similar datasets
3. The dataset is of type of data that is in high demand
4. The dataset can be used by a large number of different sectors
5. The dataset can be used in sectors where AI is expected to have a large impact on future growth
6. The dataset is of a type of data that is expected to have high commercial value in a Nordic context

Vice versa, datasets with low commercial value can be only be used in sectors where there is a small market for open data, have low existing commercial re-use, the type of data is in low demand, is expected to be used by a very limited number of different sectors and is expected to only be used in sectors where AI is expected to have a minor impact on future growth.

A criterion for the indicative market potential has also been included. The larger the geographical area covered by the dataset, the larger the potential number of users of a given AI solutions developed using the dataset.

Finally, datasets that have been collected over long periods of time and where there is a strong expectancy of continued collection in the same way have higher value for businesses that aim to be build a business model on the dataset. It is too risky for businesses to develop algorithms on datasets that might not be updated and available in the near future.

Cross-Nordic value

For datasets to create cross-Nordic value, as have been acknowledged in a European context, language barriers and interoperability aspects need to be tackled so that information resources from different organisations and countries can be combined. The availability of the information in a machine-readable format as well as a thin layer of commonly agreed metadata could facilitate data cross-reference and interoperability and therefore considerably enhance value for reuse. And the technical infrastructure needs to be in place to ensure the availability of information in the long term.[71]

Moreover, in a national context, some datasets will be too small to train efficient AI algorithms on. Volume is important where datasets are relatively generic and thus exist and display the same characteristics across the Nordic countries. There is also what could be defined as cross-border data. Some datasets are characterized by cross-national interdependencies across the Nordic countries and linking them is a prerequisite for generating value. This is true for e.g. datasets providing information about the weather, transport and traffic and data on Nordic and international trade.

The criteria measuring cross-Nordic value have been grouped into the following categories:

- Collaboration
- Dataset interoperability

The first group of criteria measures the degree to which it is possible and feasible for the Nordic countries to collaborate on a given dataset and the second group of criteria measures the potential for merging and augmenting a given dataset with data from the other Nordic countries.

Table 5 below presents the criteria for measuring cross-Nordic value.

Table 5 – Criteria for measuring cross-Nordic value

Collaboration	In how many other Nordic countries have a similar dataset been made available to the public?
	To what degree is the dataset to be characterized as cross-border (e.g. contains cross-border information)?
	Is (or can) the dataset be released so it can be found by companies and citizens in other Nordic countries?
Dataset interoperability	Is there meta-data available (in a common understood language)?
	On how many dimensions can data be linked to other Nordic datasets?
	Is merging of the dataset with similar datasets across Nordic countries necessary for sufficient dataset size and richness?

71. Open Data – An Engine for Innovation, Growth and Transparent Governance (European Commission, 2011)

Technical considerations

After applying the criteria in the assessment framework on a dataset, each of the dimension scores (A-relevance, Barriers, etc.) are normalized so they all score to 100 irrespective of the number of underlying criteria. Otherwise, AI-relevance would have a disproportionate impact on the summarized score because it contains the most criteria.

Moreover, in giving the datasets a summarized score, some dimensions are judged to be more important than others. This weight this is ultimately a political decision. In the assessments conducted for this report, the following order of importance have been used:

Estimated value for businesses

1. Cross-Nordic value / Barriers
2. Societal value
3. AI-relevance

Consequently, this e.g. means Estimated value for businesses weighs higher than Societal value; that Cross-Nordic value and Barriers are given the same importance and that AI-relevance – due to also being a selection criteria – is given the least weight.

Appendix 2: Detailed descriptions of assessed datasets

The Nordic high-value datasets assessed as a part of this project are presented with one-pagers on the following pages.

This appendix will list the assessed datasets, description of datasets, showing scores on the different value dimensions and elaborations on the scores, description and elaboration on barriers for making the datasets publicly available and/or more accessible for AI-usage.

In the following, **estimated value for businesses** has been shortened to **Business value.** It is still assessed as described in Appendix 1.

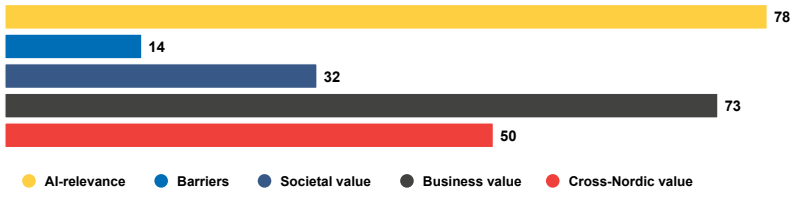
Air quality

Country	Finland
Owner of dataset (organization)	Finnish Meteorological Institute
Link to information	www.ilmatieteenlaitos.fi/ilmanlaatu
Data category	Climate / Earth observation and

This dataset provides air quality data collected from Finnish monitoring stations. The air quality map is based on modelling which combines, among others, the information about air quality measurements, weather, emissions, land use and long-range transportation. The Finnish Meteorological Institute is providing APIs for accessing the data.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 64%. The dataset is created by sensors and released without human intervention and contains several millions datapoints.
- The dataset belongs to a sector with high business value and the data has been collected in the same way for a long period of time.
- The dataset is open and accessible in 3 out of 5 of the Nordic countries. Medium added value associated with making the dataset available in all the Nordic countries.

Dataset specific barriers to openness and AI-usage

- Only access to raw data – not to the air quality model and its parameters.
- Data needs to be kept continuously updated – high demands on maintaining and updating the dataset.
- 5-10% of the dataset is composed of missing values – can induce bias in AI algorithms.

Dataset specific recommendations

- Make the air quality model and its parameters available to the public.
- Publish metadata and data descriptions to avoid bias in AI algorithms.

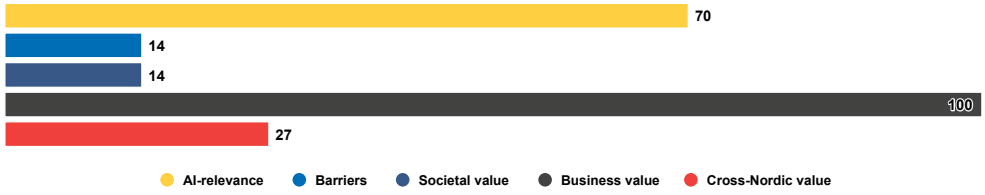
Arealressurskart – AR250 – Arealtyper

Country	Norway
Owner of dataset (organization)	Norsk institutt for bioøkonomi
Link to information	https://kartkatalog.geonorge.no/metadata/arealressurskart-ar250---arealtyper/de72929c-b250-461a-85d8-2557a2597ab4
Data category	Climate / Earth observation and

Area Management, Restriction and Regulation Zones are zones established in accordance with specific legislative requirements to deliver specific environmental objectives related to any environmental domain, for example, air, water, soil, biota (plants and animals), natural resources, land and land use.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 61%. The dataset is created by humans and updated every 3 years. The dataset is composed of geospatial maps.
- The dataset belongs to a sector with high business value and the data has been collected in the same way for a long period of time.
- The dataset is open and accessible in 3 out of 5 of the Nordic countries. Information in the dataset is very country specific.

Dataset specific barriers to openness and AI-usage

- Lack of API-access in all Nordic countries.
- Re-use cases are unclear.

Dataset specific recommendations

- Make available in all Nordic countries.
- Dataset is often used as background information to other datasets. Create visible links to these datasets to further usage and drive innovation.

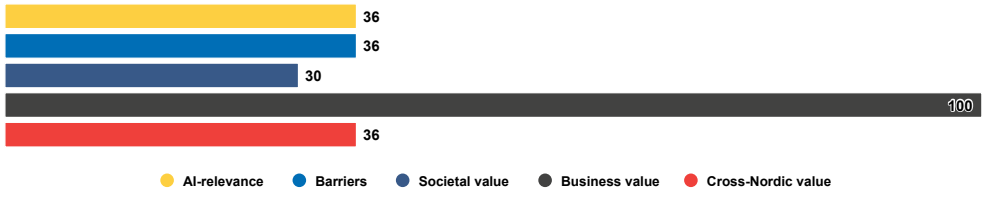
Register of bankruptcies and restructurings

Country	Finland
Owner of dataset (organization)	The Legal Register Center (Oikeusrekisterikeskus)
Link to information	www.oikeusrekisterikeskus.fi/en/index/loader.html.stx?path=/channels/public/www/ork/en/structured_nav/rekisterit/registerofbankruptciesandrestructurings_0
Data category	Companies and company specific information

The dataset contains information on bankrupties and restructurings for Finnish companies. The objective of the register is to ensure that information is made available about bankruptcy and restructuring cases. The purpose of such information is to help carry out the proceedings of courts and authorities, supervise the interests of debtors and secure the interests and rights of third parties.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 57%. The dataset is created by humans and updated regularly.
- The dataset belongs to a sector with high business value and the data has been collected in the same way for a long period of time. The dataset har high value when combined with other datasets on companies.
- The dataset is open and accessible in 1 out of 5 of the Nordic countries.

Dataset specific barriers to openness and AI-usage

- Different organizational structures in the Nordic countries mean that data is collected in different register (e.g. in a separate register in Finland and alongside general business information in Denmark).
- Some of the data is behind a paywall.

Dataset specific recommendations

- Data should be free of charge, if possible. Otherwise, costs need to be kept at a minimum.

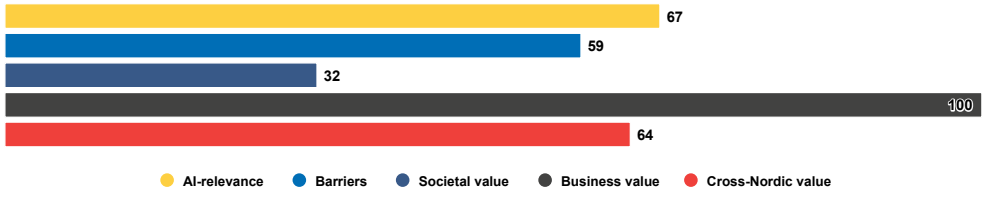
The Danish Biobank Register

Country	Denmark
Owner of dataset (organization)	Statens Serum Institut (SSI)
Link to information	www.oikeusrekisterikeskus.fi/en/index/loader.html.stx?path=/channels/public/www/ork/en/structured_nav/rekisterit/registerofbankruptciesandrestructurings_0
Data category	Health

The Danish Biobank Register collects information on samples participating in the initiative and links them to national registers, providing easy access to knowledge about available samples and number of patients with a specific diagnose. Aggregated results about the available biological material is displayed to researchers around the world through a web-based search system, to date containing information 25.3 million biological samples from 5.7 million Danes.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 62%. The dataset is created by humans and updated regularly. Data is available through a web-interface.
- The dataset belongs to a sector with high business value and the data has been collected in the same way for a long period of time. The dataset har high value when combined with other datasets on companies.
- The dataset is open and accessible in 1 out of 5 of the Nordic countries.

Dataset specific barriers to openness and AI-usage

- Raw data is highly sensitive and can only be made available as aggregates. Issues regarding anonymization needs to be fixed before data can be released.
- High initial costs in making data available, maintaining and updating it. The dataset needs to be constructed in most of the Nordic countries.
- Previous attempts at constructing a cross-Nordic biobank registry have so far not succeeded.

Dataset specific recommendations

- Even aggregated data is interesting for companies – access to the aggregated data can be facilitated.
- Show stakeholders the value added by providing visibility of data – examples available from e.g. Denmark and research.

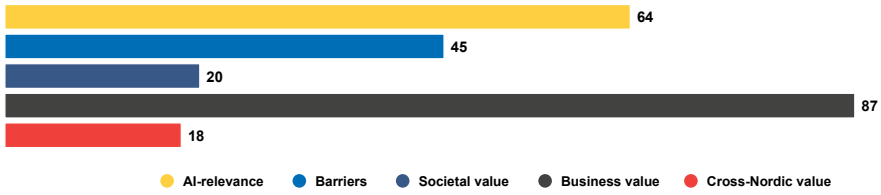
Buildings measured from flight photos

Country	Denmark
Owner of dataset (organization)	Styrelsen for Dataforsyning og Effektivisering (SDFE)
Link to information	https://sdfe.dk/saadan-arbejder-vi-med-data/flyfotos-og-laserscanning/
Data category	Climate / Earth observation and environment

The interpreted data is based on (unstructured input) flight photos and is thus values depicting the size of buildings.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 50%. The dataset is created by humans and is a solution on top of a set of raw datasets.
- The dataset belongs to a sector with high business value and the raw data has been collected for a long period of time.
- The data underlying the model is available in some of the other Nordic countries, but the model itself is not.

Dataset specific barriers to openness and AI-usage

- Proprietary solution. Not intended to be made publicly available.
- Unclear whether similar models exist in the other Nordic countries.

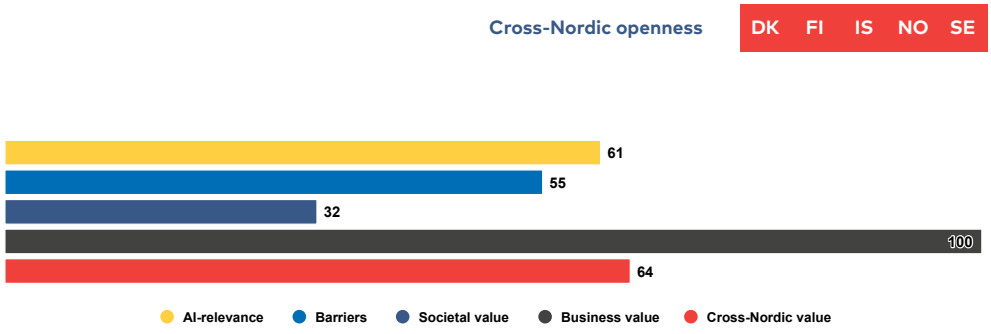
Dataset specific recommendations

- Models should be made available alongside the raw data used to build the model.
- Models should be presented on websites of public organisations, inspiring companies to what data can be used for.

Cancer Registry

Country	Iceland
Owner of dataset (organization)	Icelandic Cancer Society and the Ministry of Health (co-financing)
Link to information	www.krabb.is/krabbameinsskra/en/activities/about-icr/
Data category	Health

The Icelandic Cancer Registry (ICR) covers more than 99% of all cancer in Iceland and is a high-quality registry at the same level as the other Nordic registries. The purpose of the ICR is to gain knowledge about cancer in Iceland, to monitor the diagnosis and treatment of cancer, to ensure quality and evaluate the outcome.



Comments on assessed dataset value

- The dataset has a weighted overall score of 63%.
- The dataset is created by humans and updated regularly. The dataset is very rich and covers a population of subjects.
- The dataset belongs to a sector with high business value and the data has been collected in the same way for a long period of time.
- The dataset is only accessible for researchers and there is a high added value associated with linking data across the Nordic countries.

Dataset specific barriers to openness and AI-usage

- No access to data – data contains sensitive information on individuals.

Dataset specific recommendations

- Explore options for making dataset available in accordance with GDPR – e.g. as synthetic data.
- Provide access to aggregated data.

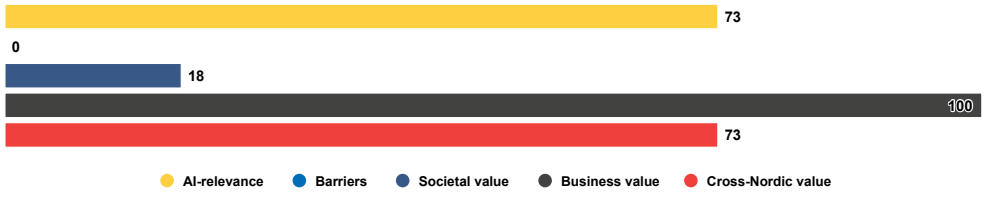
NewsWeb company announcements

Country	Norway
Owner of dataset (organization)	Oslo Børs (Oslo Stock Exchange)
Link to information	https://newsweb.oslobors.no/
Data category	Companies and company specific information

As the OAM (Official Appointed Mechanism) of Norway, Oslo Børs presents company announcements to the public.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 75%. The dataset is updated automatically when companies send in announcements to Oslo SE. The dataset contains many million observations and has been collected in the same way for a long period of time.
- The dataset is accessible for all but there is a limit to the number of requests per second. Similar datasets exist in all Nordic countries due to EU legislation.

Dataset specific barriers to openness and AI-usage

- Copyright issues.
- Detailed information can be behind paywall or only made available to certain companies.
- Not necessarily *public* data in all Nordic countries when the OAM is a private company (e.g. NASDAQ)

Dataset specific recommendations

- Data needs to be licensed differently to ease access for companies. A legal study of whether or not the data needs to be copyrighted can be conducted.
- Data needs to be available free of charge.
- Contact needs to be taken to the private owners of the data – is it possible to make data more accessible? Public organizations can offer to host data so the private company does not incur a cost.

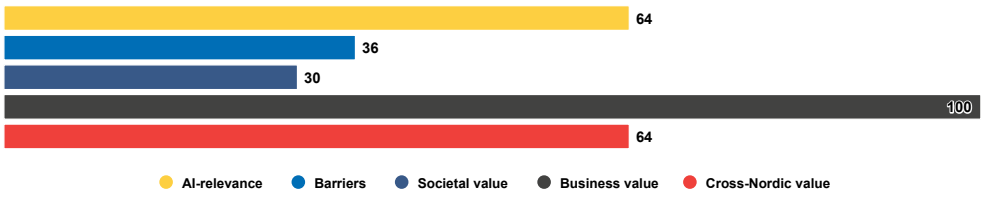
National business registry

Country	Sweden
Owner of dataset (organization)	Bolagsverket (Swedish Companies Registration Office)
Link to information	https://bolagsverket.se/en/us/about/e-services/foretagsfakta
Data category	Companies and company specific information

Information on Swedish business, e.g. registration certificates or annual reports.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 66%. The dataset contains many observations and many variables linked to companies. The dataset has been collected in the same way for a long period of time.
- The dataset belongs to a sector with high business value. The dataset has high value when combined with other datasets on companies. The dataset is open and easily accessible in 2 out of 5 of the Nordic countries.

Dataset specific barriers to openness and AI-usage

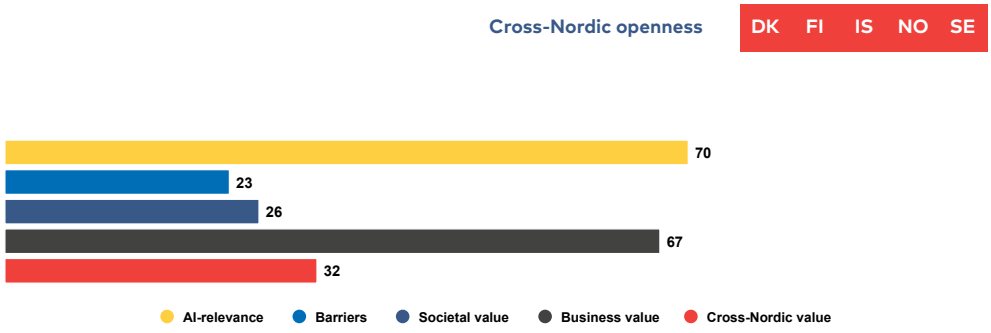
- Some of the data is behind a paywall.
- Similar datasets do not contain the same information across the Nordic countries.

Dataset specific recommendations

- The Nordic Smart Government programme is currently working with the five national business registers. Any work related to making this data publicly available should take place in the context of the NSG-programme.

Data on product tests

Country	Denmark
Owner of dataset (organization)	Sikkerhedsstyrelsen (Danish Safety Technology Authority)
Link to information	
Data category	Public governance
Dataset with information on product tests conducted by the Danish Safety Technology Authority.	



Comments on assessed dataset value

- The dataset has a weighted overall score of 54%. The dataset is created by humans and data contains both value in columns and rows, and text and images. The dataset has been collected in the same way for a long period of time.
- The dataset belongs to a sector with medium business value.
- The dataset is difficult or impossible to locate in all the Nordic countries.

Dataset specific barriers to openness and AI-usage

- Datasets are not findable by companies or viewable only.

Dataset specific recommendations

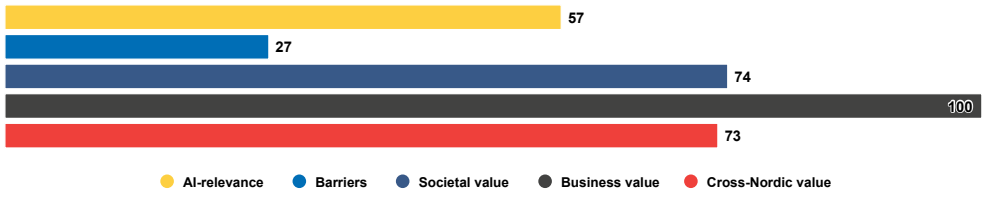
- Data should be made visible for companies.
- Data should be made available as a download.

Energi Data Service (Platform for accessing various energy-related datasets)

Country	Denmark
Owner of dataset (organization)	Energinet
Link to information	
Data category	Climate / Earth observation and environment
Data about the Danish energy system such as CO2 emissions and consumption and production data.	

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 78%. The dataset is created by sensors by require human interaction before release. The dataset has been collected in the same way for a long period of time.
- The dataset belongs to a sector with high business value.
- The dataset is available as aggregates in all of the Nordic countries. Raw data is sensitive and has not been made publicly available.

Dataset specific barriers to openness and AI-usage

- No access to data – data contains sensitive information on individuals.
- Making data available is costly and requires extensive data management.

Dataset specific recommendations

- Explore options for making dataset available in accordance with GDPR – e.g. as synthetic data.

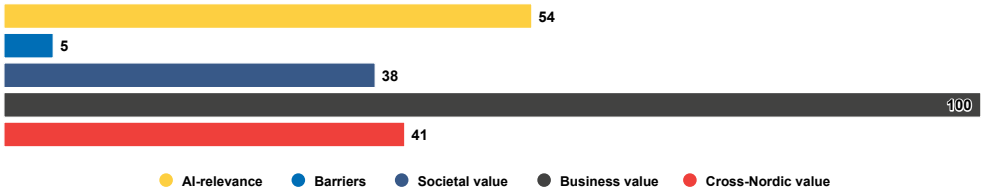
Hydrological interface

Country	Finland
Owner of dataset (organization)	Finnish Environment Institute (SYKE)
Link to information	www.avoindata.fi/data/en_GB/dataset/hydrologi arajapinta / http://metatieto.ymparisto.fi:8080/g eoportal/catalog/search/resource/details.page?uu id=%7B86FC3188-6796-4C79-AC58-8DBC7B5688 27%7D
Data category	Climate / Earth observation and environment

Information on the regional and temporal distribution of water resources in Finland is published through the hydrological interface. Observations are made on the elements of the hydrological cycle (precipitation, evaporation, flow and runoff), the amount of water (water level in water bodies) and other hydrological phenomena (water value of snow, ice thickness, water temperature, etc.).

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 69%. The dataset is mostly created by sensors. The dataset has been collected in the same way for a long period of time.
- The dataset belongs to a sector with high business value.
- The dataset is available in most of the Nordic countries. Not all information is equally relevant in all Nordic countries due to different geographies (e.g. ice thickness information).

Dataset specific barriers to openness and AI-usage

- Swedish dataset has global coverage but paywalled for bigger datasets.

Dataset specific recommendations

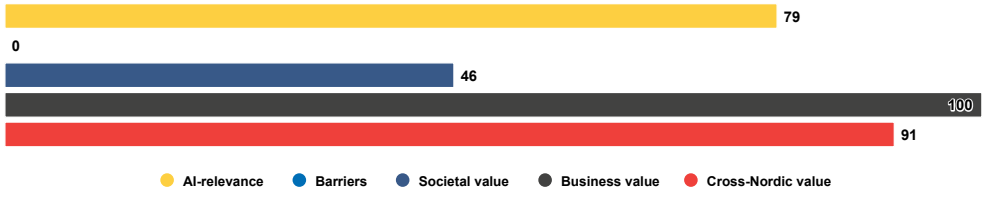
- Make data available free of charge.
- Ensure interoperability of cross-Nordic datasets incl. language of dataset descriptions and metadata

Groundwater

Country	Sweden
Owner of dataset (organization)	The Geological Survey of Sweden (SGU)
Link to information	www.sgu.se/produkter/geologiska-data/oppna-data/grundvatten-oppna-data/
Data category	Climate / Earth observation and environment
Many aspects of groundwater, location, time series, historical, sources, environmental monitoring and water quality.	

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 85%. The dataset is large and rich. The dataset has been collected in the same way for a long period of time.
- The dataset belongs to a sector with high business value.
- The dataset is available in all the Nordic countries. Dataset value will increase if linked across the Nordic countries.

Dataset specific barriers to openness and AI-usage

- Data is not available through APIs in all Nordic countries.

Dataset specific recommendations

- Make data available through APIs.
- Ensure interoperability of cross-Nordic datasets incl. language of dataset descriptions and metadata

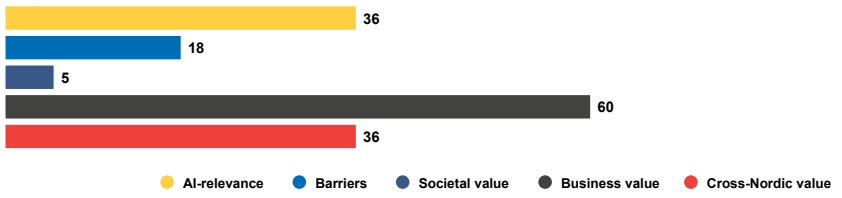
Database of types of nature

Country	Norway
Owner of dataset (organization)	Artsdatabanken (Institution under Kunnskapsdepartmentet)
Link to information	www.naturtyper.artsdatabanken.no/
Data category	Climate / Earth observation and environment

The dataset describes how the Norwegian nature can be classified and separated in different types of nature.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 46%. The dataset highly unstructured with images and text on different webpages. The dataset has NOT been collected in the same way for a long period of time and it is unclear if it is being updated regularly.
- The dataset belongs to a sector with high business value.
- Dataset either does not exist or is not made publicly available in most of the Nordic countries.

Dataset specific barriers to openness and AI-usage

- Data is not easily accessible for machines – data is presented on a set of webpages.

Dataset specific recommendations

- Make data easily downloadable.
- Create visibility around Danish use-case of dataset (nature recognition) to promote use and re-use of data.

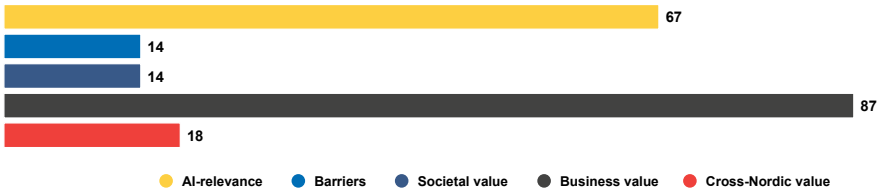
Regulation plans

Country	Norway
Owner of dataset (organization)	Kommunerne (Norwegian Municipalities)
Link to information	https://fellesdatakatalog.digdir.no/datasets/0415872e-3fa7-48e0-aa8e-ec90ab47d27d%20/
Data category	Public governance

Regulation plans are area maps that determine the use and protection of specific areas and that provide the basis for decisions about building and planning in those areas. Developed by the municipalities.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 56%. The dataset is composed of geospatial map layers. The dataset has been collected in the same way for a long period of time. The dataset belongs to a sector with medium business value.
- Private users need to contact the municipality in question in order to request data. Unclear if historic data has been digitalized in the Nordic countries.

Dataset specific barriers to openness and AI-usage

- Data is not easily accessible in a central place.
- Many different data stakeholders.

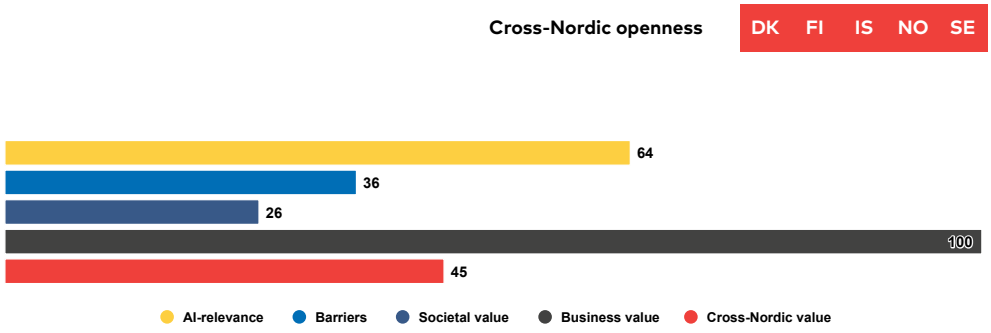
Dataset specific recommendations

- Make data easily downloadable in a central location. Alternatively, link to municipality webpages where data is accessible.

Danish Rheumatological Database (DANBIO)

Country	Denmark
Owner of dataset (organization)	DANBIO, funded by Danish Regions
Link to information	www.rkkp-dokumentation.dk/Public/Databases.aspx?db=26&version=3
Data category	Health

The Danish Rheumatology Database (DANBIO) is a nationwide clinical quality database that gathers data on patients with arthritis and being treated with biological drugs for rheumatic disease in Denmark. The aim of DANBIO is to ensure effective treatment of individual patients while the collated data is valuable in scientific studies.



Comments on assessed dataset value

- The dataset has a weighted overall score of 62%. The dataset is collected by humans but contain ground truths on patients with rheumatic diseases.
- The dataset has been collected in the same way for a long period of time and belongs to a sector with high business value. The dataset is not available across the Nordic countries and it is highly doubtful that similar datasets in the other Nordic countries contain comparable information (e.g. same variables).

Dataset specific barriers to openness and AI-usage

- Data is not accessible for non-researchers.
- Data contains sensitive information on individuals.

Dataset specific recommendations

- Explore options for making dataset available in accordance with GDPR – e.g. as synthetic data.
- Provide access to aggregated data.
- If similar datasets exist across the Nordic countries, ensure interoperability through e.g. same set of variables.

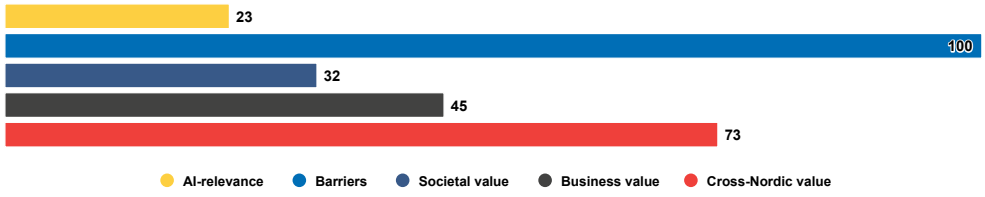
Swedish CArdioPulmonary bioImage Study (SCAPIS)

Country	Sweden
Owner of dataset (organization)	SCAPIS National Steering Committee (Research Collaboration)
Link to information	http://scapis.org/
Data category	Health

SCAPIS is a large research study aimed at predicting and preventing cardiovascular and chronic obstructive pulmonary disease. The goal is to further develop individualised treatment and improve health care by building a nationwide, open-access, population-based cohort. SCAPIS has recruited and investigated 30,000 men and women aged 50 to 64 years with detailed imaging and functional analyses of the cardiovascular and pulmonary systems. Data is geotagged, extensive and continuously growing.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 57%. The dataset is collected by humans but contains ground truths on patients with cardiovascular diseases. The data contains images.
- The dataset is newly collected and belongs to a sector with high business value. The dataset is not available across the Nordic countries and it is highly doubtful that similar datasets in the other Nordic countries contain comparable information (e.g. same variables).

Dataset specific barriers to openness and AI-usage

- Data is not accessible for non-researchers.
- Data contains sensitive information on individuals.
- Data is unique to Sweden.
- Data is expensive to collect and maintain.

Dataset specific recommendations

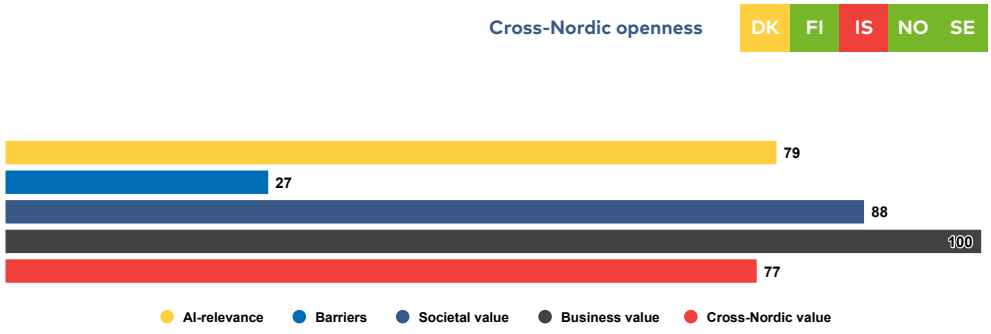
- Explore options for making dataset available in accordance with GDPR – e.g. as

synthetic data.

- Provide access to aggregated data.
- If similar datasets exist across the Nordic countries, ensure interoperability through e.g. same set of variables.

Swedish Meteorological and Hydrological data

Country	Sweden
Owner of dataset (organization)	SMHI
Link to information	www.smhi.se/data/utforskaren-oppna-data/
Data category	Climate / Earth observation and environment
Online (near real-time) and historical data of weather forecasts and observations.	



Comments on assessed dataset value

- The dataset has a weighted overall score of 84%. The dataset is collected by sensors and contains datasets augmented with predictive algorithms (e.g. weather forecasts).
- The dataset has been collected in the same way for a long period of time and belongs to a sector with high business value. The dataset is available in 3 of the 5 Nordic countries. The dataset can be characterized as a cross-border dataset, where observations in one country impact observations in another country.

Dataset specific barriers to openness and AI-usage

- There are already projects underway for making these datasets available in the countries where it has not been done (e.g. the Danish Meteorological Institute has received funding and compensation for the lack of income – data is expected to become available in the period 2020-2022).

Dataset specific recommendations

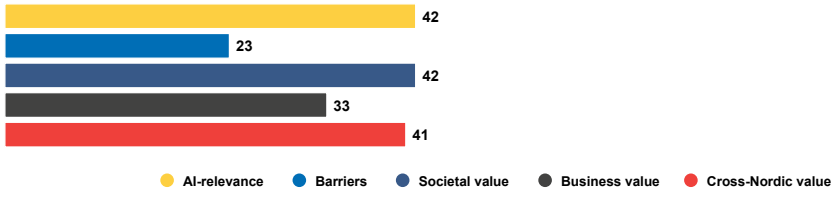
- There are already projects underway for making these datasets available in the countries where it has not been done (e.g. the Danish Meteorological Institute has received funding and compensation for the lack of income – data is expected to become available in the period 2020-2022).

Spoken Language

Country	Iceland
Owner of dataset (organization)	Reykjavík University and The Icelandic Centre for Language Technology
Link to information	www.malfong.is/index.php?lang=en&pg=malromur
Data category	Culture
The Malromur corpus is an open source corpus of Icelandic voice samples.	

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 47%. The dataset contains between 100,000 and 1,000,000 audio datapoints from voice samples in combination with text samples. The dataset has a clear owner structure and is freely available.
- The dataset is regional, and the data is a one-time collection. The dataset belongs to a sector with low business value, and similar datasets have not been made available in the other Nordic countries.

Dataset specific barriers to openness and AI-usage

- Datasets are only available to researchers. Reluctancy to make datasets available to everyone.
- Datasets cover very specific populations and are only collected as part of research projects.
- Datasets contain sensitive information.

Dataset specific recommendations

- Make datasets publicly available. There are examples across the Nordic countries of research data being made publicly available.
- Ensure good data descriptions and metadata.
- Publish data without identifiers.

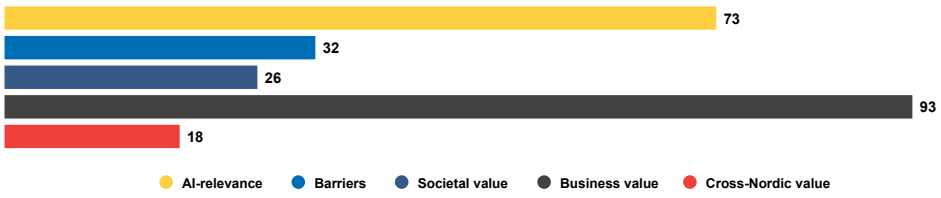
Surface water

Country	Norway
Owner of dataset (organization)	Kartverket (Norwegian Mapping Authority)
Link to information	https://kartkatalog.geonorge.no/metadata/595e47d9-d201-479c-a77d-cbc1f573a76b
Data category	Climate / Earth observation and environment

The dataset (FKB-Vann) describes geographic location, the shape and the course of surface water flows in Norway.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

The dataset has a weighted overall score of 56%. The dataset has more than 1 million datapoints, but data is collected and released manually. Data is almost complete and has been collected in many time periods. The dataset is available through API, and it generates revenues by being sold. The data has been collected for more than 3 years and is generated at country level. Metadata exist in a common understood language, but the dataset Is national, and the potential for linking with other datasets is low.

Dataset specific barriers to openness and AI-usage

- The dataset generates value by being sold.

Dataset specific recommendations

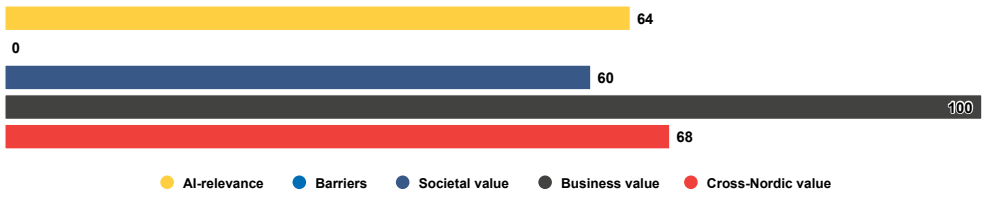
- The dataset owner needs to be compensated for the revenue lost.

Traffic events and roadworks

Country	Iceland
Owner of dataset (organization)	Icelandic Road and Coastal Administration
Link to information	http://www.road.is/travel-info/road-conditions-and-weather/south-iceland-road-conditions-map/
Data category	Mobility
Information on traffic events (accidents etc.) and roadworks on Icelandic roads.	

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

The dataset has a weighted overall score of 81%. The dataset, which is released as map data, has between 10,000 and 100,000 datapoints from several time dimensions with GPS coordinates. There is a clear owner structure, no GDPR-inflictions, low cost associated with making the dataset public, and possibility of releasing the dataset through API. The data has been collected for more than 3 years, and it is in a sector with high business value. Mobility is a cross-Nordic issue, data is available for all, and there is potential for linking it across countries and other datasets.

Dataset specific barriers to openness and AI-usage

- IT-infrastructure needs to be updated in order to facilitate access to data. Data is currently presented online but cannot be downloaded.

Dataset specific recommendations

- Enable data download, preferably through an API.
- Publish data through the national open data portal.

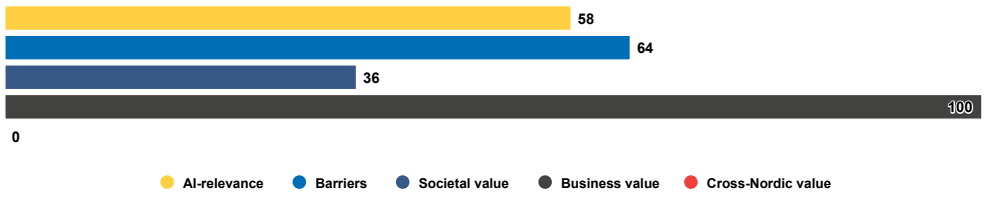
The Danish National Waste Register

Country	Denmark
Owner of dataset (organization)	The Danish Environmental Protection Agency
Link to information	https://mst.dk/affald-jord/affald/affaldsdatasystemet/
Data category	Earth observation and environment

Information about waste production per municipality, year, type of waste, source and treatment and re-use percentages

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

- The dataset has a weighted overall score of 47%. The dataset has more than one million datapoints collected and released manually. The data does however not contain authoritative truths and has only one-time dimension. It is clear who is responsible for the dataset, and it cannot currently be bought. The data holds high business value, and it has been collected for more than three years.
- There is limited cross Nordic value as it is data owned by municipalities, with few links to other datasets, and it cannot be made available as it contains sensitive information.

Dataset specific barriers to openness and AI-usage

- Data contains sensitive information on companies.
- The update frequency of data is very low – large lag from collection to database.
- Existence of similar datasets across the Nordic countries is unclear.

Dataset specific recommendations

- Aggregate data to a degree where it is no longer sensitive and publish through the national open data platform.
- Ensure better collection of data so that data is updated more frequently.

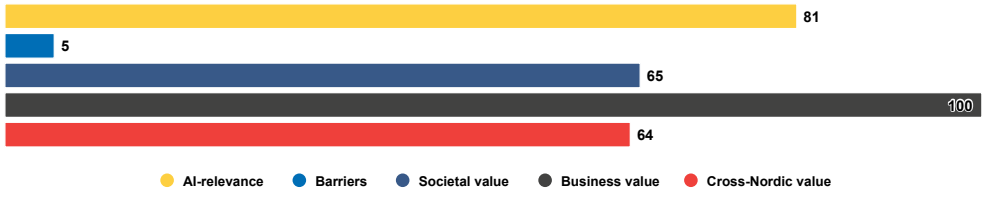
Road web cameras

Country	Finland
Owner of dataset (organization)	Traffic Management Finland
Link to information	https://vayla.fi/web/en/open-data/digitraffic
Data category	Mobility

API photostream from webcameras located alongside the Finnish roads. Cameras provide information on current traffic flow and weather conditions.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

The dataset has a weighted overall score of 82%. The dataset high AI-relevance with more than one million datapoints generated by sensors. Furthermore, the data has GPS coordinates and contain several time dimensions. There are few barriers, but there is some cost associated with setting up an API incl. storing historic data. The dataset has high business value, aa the category is mobility and data has been collected for more than three years. The cross Nordic value is based on a high degree of cross-border data, high availability, metadata and potential for merging the dataset with other datasets.

Dataset specific barriers to openness and AI-usage

- Dataset is open and available through an API. Work needs to be done in the other Nordic countries

Dataset specific recommendations

- Dataset is open and available through an API. Work needs to be done in the other Nordic countries

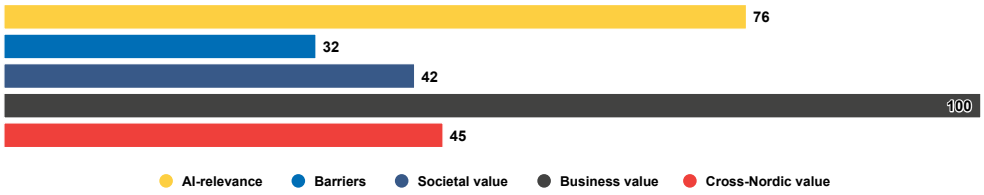
Occupational accident report register

Country	Finland
Owner of dataset (organization)	Ministry of Social Affairs and Health
Link to information	https://www.stat.fi/til/ttap/2017/ttap_2017_2019-11-29_tie_001_en.html
Data category	Health

The occupational accidents register contains information on which occupational accidents have happened, when, where and what kind of accident. It is aggregated in several dimensions, e.g. type of employment, sector and severity.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

The dataset has a weighted overall score of 67%. The high AI-score is due to more than one million datapoints (which grows with 130,000 each year), certification labels, and several time dimensions. The barriers are GDPR-related, unclear responsibility for misuse, and low to medium cost of releasing and maintaining the dataset. The business value is high due to a high value sector, a national scope and structured data collection for more than three years. The cross Nordic value is limited by uncertainty on whether similar datasets can be found in the other Nordic countries.

Dataset specific barriers to openness and AI-usage

- Raw data cannot be published due to sensitive data on individuals. Aggregated data might not be aggregated according to business' needs.
- Aggregated datasets can be viewed but not downloaded easily.

Dataset specific recommendations

- Enter dialogue with business for which aggregations are most relevant for businesses.
- Create visibility of dataset through national open data platform.
- Enable option for downloading data.

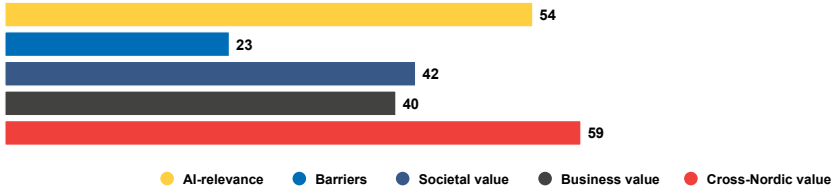
ParIce (an English-Icelandic Parallel Corpus)

Country	Island
Owner of dataset (organization)	Unclear
Link to information	www.malfong.is/index.php?lang=en&pg=samhlida
Data category	Culture

This is the first parallel corpus built for the purposes of language technology development and research for Icelandic, although some Icelandic texts can be found in various other multilingual parallel corpora.

Cross-Nordic openness DK FI IS NO SE

Assessed dataset value



Comments on assessed dataset value

The dataset has a weighted overall score of 54%. Data contains unstructured text, ground truths. Important barriers are high costs of preparing for release and maintenance, lack of API, and unclear responsibility for misuse. Data has been gathered for more than three years, but in a sector that does not hold much commercial value. The cross Nordic value is based on Nordic availability and cross border relevance.

Dataset specific barriers to openness and AI-usage

- Datasets are only available to researchers. Reluctancy to make datasets available to everyone.
- Datasets cover very specific populations and are only collected as part of research projects.
- Datasets contain sensitive information.

Dataset specific recommendations

- Make datasets publicly available. There are examples across the Nordic countries of research data being made publicly available.
- Ensure good data descriptions and metadata.
- Publish data without identifiers.

Consulted organizations

Representatives and data-owners from the following organizations have provided input at different stages in the project. We would like to direct thanks to all of them for invaluable insights.

Denmark
The Danish Biobank Register
The Danish Business Authority
The Danish Environmental Protection Agency
The Danish Road Directorate
Energinet

Finland
Digital and Population Data Services Agency
Finland's Environmental Administration
Finnish Institute for Health and Welfare
Finnish Meteorological Institute

Iceland
The Environment Agency of Iceland
Icelandic Road and Coastal Administration

Norway
Kartverket
Norwegian Biodiversity Information Centre
Norwegian Institute for Air Research
Oslo Stock Exchange

Sweden
DIGG - Agency for Digital Government
Ignite Sweden
Lantmäteriet
Nordic Innovation of Sweden
RISE
Samtrafiken
Swedish Meteorological and Hydrological Institute

Other
NordForsk
NordicAI

About this publication

Nordic cooperation on data to boost the development of solutions with artificial intelligence

Nord 2020:042
ISBN 978-92-893-6674-8 (PDF)
ISBN 978-92-893-6675-5 (ONLINE)
http://doi.org/10.6027/nord2020-042

Layout: Gitte Wejnold

Nordic co-operation

Nordic co-operation is one of the world's most extensive forms of regional collaboration, involving Denmark, Finland, Iceland, Norway, Sweden, and the Faroe Islands, Greenland and Åland.

Nordic co-operation has firm traditions in politics, economics and culture and plays an important role in European and international forums. The Nordic community strives for a strong Nordic Region in a strong Europe.

Nordic co-operation promotes regional interests and values in a global world. The values shared by the Nordic countries help make the region one of the most innovative and competitive in the world.

The Nordic Council of Ministers
Nordens Hus
Ved Stranden 18
DK-1061 Copenhagen
pub@norden.org

Read more Nordic publications on www.norden.org/publications