

FINAL REPORT

Territorial indicators based on big data

Main Report // November 2021

This Final Report is conducted within the framework of the ESPON 2020 Cooperation Programme, partly financed by the European Regional Development Fund.

The ESPON EGTC is the Single Beneficiary of the ESPON 2020 Cooperation Programme. The Single Operation within the programme is implemented by the ESPON EGTC and co-financed by the European Regional Development Fund, the EU Member States, the United Kingdom and the Partner States, Iceland, Liechtenstein, Norway and Switzerland.

This delivery does not necessarily reflect the opinions of members of the ESPON 2020 Monitoring Committee.

Authors

Prognos AG: Dr. Jan-Philipp Kramer, Dr. Georg Klose, Johanna Thierstein, Lennart Galdiga, Janis Neufeld

DevStat: Dr. José Vila, José L. Cervera, Yolanda Gòmez

External expert: Prof. Dr. Dirk Heckmann (TU München)

ESPON EGTC: Zintis Hermansons (project expert) and Caroline Clause (financial expert)

Information on ESPON and its projects can be found at www.espon.eu.

The website provides the possibility to download and examine the most recent documents produced by finalised and ongoing ESPON projects.

ISBN: 978-2-919816-24-8

© **ESPON, 2020**

Layout and graphic design by BGRAPHIC, Denmark

Printing, reproduction or quotation is authorised provided the source is acknowledged and a copy is forwarded to the ESPON EGTC in Luxembourg.

Contact: info@espon.eu

FINAL REPORT

Territorial indicators based on big data

Main Report // November 2021

Table of contents

Abbreviations	7
1 Introduction	8
1.1 Background and objectives of the study	8
1.2 Overall approach of the study	10
1.3 Structure of the report	10
2 Overview of relevant private digital platforms for territorial analysis	12
2.1 Prequalified long-list of private digital platforms	12
2.2 Final list of private digital platforms	14
2.2.1 Criteria for the short list	14
2.2.2 Qualitative and quantitative analysis by type of platform	15
2.2.3 Short List of selected platforms	20
3 Framework: Analysis of the legal aspects of access to data of selected platforms	21
3.1 General information on accessing data of platforms	21
3.1.1 Entitlement to data access	21
3.1.2 "Contract resolution": access only to voluntarily provided data sets	22
3.1.3 Restrictions on data access and data use	23
3.1.4 Key legal aspects for data access	25
3.2 Analysing the specific requirements of data access and processing of the selected platforms	26
3.2.1 Components of the framework for for accessing data from private digital platforms	26
3.2.2 Twitter	27
3.2.3 LinkedIn	30
3.2.4 Google Maps	31
3.2.5 TripAdvisor	33
3.2.6 ResearchGate	34
3.2.7 Key takeaways for the framework of cooperation	34
4 Territorial indicators based on big data	36
4.1 Georeferencing of territorial indicators	36
4.2 Estimation strategies for territorial indicators based on data from private digital platforms	36
4.3 Reference period and breakdown of territorial indicators	38
4.4 Overview of potential territorial indicators	39
4.5 Selected indicators for the demonstrator study	42
4.6 Going beyond the basic territorial indicators: Second-tier indicators	49
5 Demonstrator Study: Territorial big data	51
5.1 Selection of pilot regions	51
5.2 LinkedIn: Technical skills for digital innovation	52
5.2.1 Methodology	52
5.2.2 Illustrations of findings on technical skills for digital innovation	54
5.3 Twitter: COVID-19	58
5.3.1 Methodology	58
5.3.2 Illustrations of findings on Covid-19	60
6 Concluding remarks and outlook	64
7 Annex	66
7.1 Long-list of selected platforms	66
7.2 Results / findings from Focus Group 1 (27 th May 2021)	72
7.3 Results / findings from Focus Group 2 (1 st July 2021)	74
7.4 Experts that participated in the focus groups	76
References	77

List of figures, tables and boxes

List of figures

Figure 1-1: Big data triangle on territorial big data	9
Figure 1-2: Overview of the project approach	10
Figure 2-1: Territorial Agenda 2030 – a reference point for this project	13
Figure 2-2: Schematic overview over the qualification process	13
Figure 2-3: 20 pre-selected private platforms	14
Figure 3-1: Component of the framework for cooperation for the analysis of data access & data usage.....	26
Figure 3-2: Relevant information sources	27
Figure 5-1: Share of LinkedIn Members in the employed population	54
Figure 5-2: Share of LinkedIn Members with skills in the field of Programming, Robotics, and AI in the total LinkedIn Members by country.....	55
Figure 5-3: Share of LinkedIn Members with skills in the field of Programming, Robotics, and AI in the active population by country	56
Figure 5-4: Share of LinkedIn Members with skills in the field of AI and Robotics in the total LinkedIn Members by region.....	56
Figure 5-5: Compare the share of LinkedIn Members with programming skills in the employed population on NUTS 3 level with employment in high-tech sectors by NUTS 2 region.....	57
Figure 5-6: Number of tweets on Covid-19 per label (subtopic)	60
Figure 5-7: Number of tweets on Covid-19 per 10,000 capita.....	60
Figure 5-8: Number of infections and COVID19 salience	63
Figure 5-9: <i>Vaccination rates and COVID19 salience</i>	63
Figure 7-1: Results regarding challenges and difficulties in accessing big data from private digital platforms.....	72
Figure 7-2: Experiences with platforms from the prequalified long list	73
Figure 7-3: Experiences regarding specific cooperation agreements with private digital platforms	74
Figure 7-4: What specific indicators could be drawn from big data? - Twitter.....	74
Figure 7-5: What specific indicators could be drawn from big data? - LinkedIn	75
Figure 7-6: What specific indicators could be drawn from big data? - TripAdvisor.....	75
Figure 7-7: What specific indicators could be drawn from big data? - ResearchGate.....	76

List of tables

Table 1-1: Structure of the Final Report.....	11
Table 2-1: Platform profiles by type of platform.....	16
Table 2-2: Quantitative analysis of micro blogging platforms	18
Table 2-3: Quantitative analysis of social media platforms	19
Table 2-4: Quantitative analysis of platforms regarding tourism	19
Table 2-5: Quantitative analysis of knowledge platforms	20
Table 2-6: Summary of selected platforms	20
Table 3-1: Overview of Twitter data access options.....	27
Table 3-2: Factsheet Twitter API.....	28
Table 3-3: Overview of LinkedIn data access options.....	30
Table 3-4: Factsheet LinkedIn API.....	31
Table 3-5: Overview of Google Maps data access.....	32
Table 3-6: Factsheet Google Maps API	32
Table 3-7: Overview of TripAdvisor data access.....	33
Table 3-8: Factsheet Tripadvisor API.....	33
Table 3-9: Overview of ResearchGate data access.....	34
Table 3-10: Overview of API availability and restrictions	35

Table 3-11: Elements of the data access framework	35
Table 4-1: Information on Twitter's user provided by the API	39
Table 4-2: Overview of potential territorial indicators (selected examples, non-exhaustive)	40
Table 4-3: Indicators relevant to the perception of COVID-19 crisis	43
Table 4-4: Indicators relevant for technical skills for digital innovation	44
Table 4-5: Indicators relevant for quality of life.....	46
Table 4-6: Indicators relevant to access to bank services.....	47
Table 4-7: Indicators relevant to land use	48
Table 4-8: Indicators relevant to R&D employment territorial indicators	49
Table 5-1: Selected Regions for pilot study	51
Table 5-2: Ontology for selected technology areas.....	53
Table 5-3: Ontology on the Topic Covid 19.....	59
Table 5-4: Infections, vaccination, and salience of COVID19	62
Table 7-1: Full long list of 79 digital platforms providing big data.....	66
Table 7-2: Experts that participated in the focus groups	76

List of boxes

Box 3-1: Application Programming Interface - definition	23
Box 3-2: Twitter Academic Research Track	29
Box 3-3: LinkedIn Economic Graph	31
Box 4-1: Estimation with distances	37
Box 4-2: Generic indicators.....	42
Box 4-3: Spatial data analysis.....	50

Abbreviations

API	Application Programming Interface
DA	Daily Actives
DGINS	Directors General of the EU National Statistical offices
EC	European Commission
EFTA	European Free Trade Association
ESPON	European Territorial Observatory Network
ESS	European Statistical System
ESSnet	European Statistical System Network
LFS	Labour Force Survey
LORDI	Local and Regional Digital Indicators
MA	Monthly Actives
NGO	Non-profit organization
NUTS	Nomenclature of Territorial Units for Statistics
SDGs	Sustainable Development Goals

1 Introduction

1.1 Background and objectives of the study

The key objectives of this project are to provide a **qualified list of private platforms** which can provide comparable data for a European wide territorial analysis and **to work towards a framework of cooperation that would allow for lasting data extraction possibilities.**

The **rationale** behind this stems from the social benefits that this data can bring to many areas, for instance regarding the spatial distribution of property prices or patterns of research and innovation. Moreover, the intelligent use and analysis of these data volumes is increasingly becoming a critical competitive parameter for companies and public authorities alike. Big data technologies and applications can unlock the potential of these increasing data volumes and analysis requirements for decision-makers in industry and policy and make them usable.

However, **the use of big data to inform public policy decision-making is still scarce.** There have been some projects that build upon the advantages of big data, such as being timelier, more context-specific, and more spatially precise compared to official statistics. Important projects in this respect are the ESSnet Big Data I (2016-18) and II (2018-2020) projects, which covered subject matters such as web scraping of job vacancies, web scraping of enterprise characteristics, mobile phone data, or early estimates. Similarly, the UN's Global Working Group on 'Big Data for Official Statistics' is to be highlighted. It collaborates with the Google Earth Engine to Monitor SDGs. The European Commission has also worked on important projects to further facilitate the use of big data in public policy work, including the report "Towards a European strategy on business-to-government data sharing for the public interest"¹ or the analytical report "Business-to-Government Data Sharing"². In the field of European official statistics, at the Directors General of the EU National Statistical offices' (DGINS) meeting of September 2013, the heads of EU statistical offices recognised, as stated in the Scheveningen Memorandum³, the relevance of big data for the ESS and the need for adopting a related action plan considering the development of methodology, capabilities, and a legislative framework, to be implemented in partnership with governments, academics, and private sources.⁴ In this context, the **European Commission has also prominently finalised an agreement between the tourism platforms Booking.com, TripAdvisor, Airbnb, and Expedia to share data with Eurostat in January 2020** to get a better view on developments regarding holidays accommodations.⁵ Based on this agreement Eurostat has recently published experimental statistical data on the number of guest nights spent at short-stay accommodations at the NUTS 2 level in the EU.⁶

The **role of spatial data** is of key importance in this context. As outlined above, one of the key advantages of big data is that it is often spatially more precise than official statistics. There is a growing number of scientific research on big spatial data regarding applicable spatial analysis methods and tools such as research by Yoshiki Yamagata and Hajime Seya (2020) on "Spatial Analysis Using Big Data"⁷ which discusses how high-quality real-time data could be used in order to analyse socio-economic developments in cities or

¹ https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=64954

² https://www.europeandataportal.eu/sites/default/files/analytical_report_12_business_government_data_sharing.pdf

³ <https://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13>

⁴ In fact, Eurostat set up an internal Task Force on Big Data with the lifespan 2014-2016 with the following objectives: to lead and co-ordinate developments within the ESS and the European Commission with regard to maximising the potential of Big Data for Official Statistics and evidence-based policy making and to develop — together with all members of the ESS — an ESS Big Data strategy along the lines of the Scheveningen Memorandum.

⁵ https://ec.europa.eu/commission/presscorner/detail/en/ip_20_194

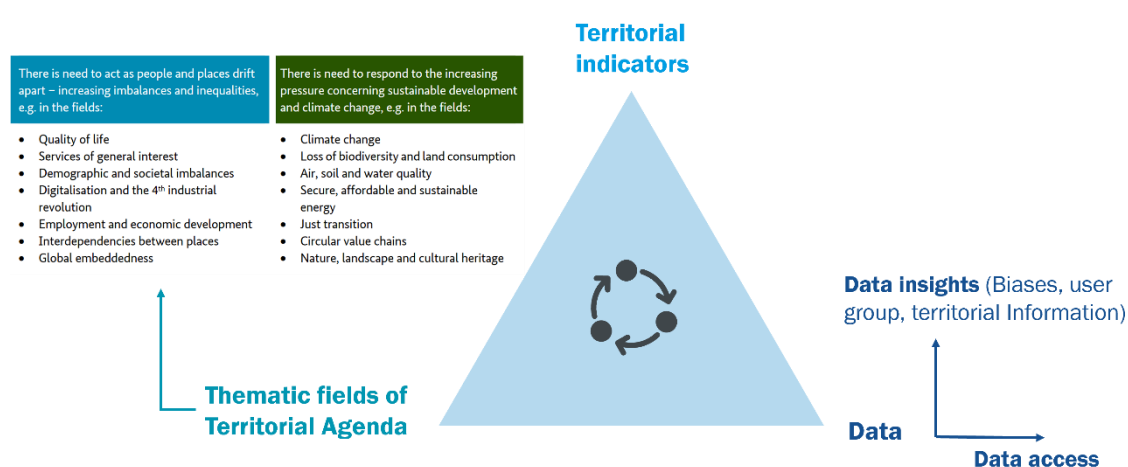
⁶ https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Short-stay_accommodation_offered_via_online_collaborative_economy_platforms

⁷ <https://www.sciencedirect.com/book/9780128131275/spatial-analysis-using-big-data?via=ihub=>

work by Angioletta Voghera and Luigi La Riccia from Politecnico di Torino on Spatial Planning in the Big Data Revolution which discusses, amongst others, how big data volumes can be connected to GIS to perform spatial evaluation tools. Other examples, among others from ESPON projects as well as other research (use of Twitter & Facebook data; Google maps data; Flickr data on geotagged photos, LinkedIn & World Bank on skills and talent migration; etc.), formed an important starting point for this project.⁸ Moreover, ESPON is currently working on the project 'LORDI' (Local and Regional Digital Indicators), for which it has engaged in big data extraction with more than 20 platforms. The learnings from this endeavour were taken into account for this project as well.

Generally, this endeavour does not have any specific sectoral coverage, rather it aims to be in line with the topics outlined in the **Territorial Agenda 2030**⁹ and strives to describe territorial development patterns and trends in terms of disparities, peripheralization, convergence, urbanization, territorial decentralization, mobility and/or other territorial processes. The described interlinkage between territorial indicators, big data, and thematic fields ('big data triangle') that is at the core of this project is summarised in the following Figure 1-1.

Figure 1-1: Big data triangle on territorial big data



Source: Prognos AG/DevStat (2021).

In the context of this approach, **this project aims to achieve the following outcomes:**

1. To provide a **list of private digital platforms** that could be valuable big data providers for a European wide territorial analysis.
2. To design a **framework of cooperation** with private digital platforms, which would allow big data sharing and extraction on a regular basis to develop territorial indicators for public policy analysis.
3. To come to a **list of territorial indicators** which help to use the provided big data along with the information on how to interpret the data behind the indicators.
4. To provide **brief descriptions on territorial trends** for a sample of the selected indicators, alongside graphic illustrations.

An overview of the project approach is provided below.

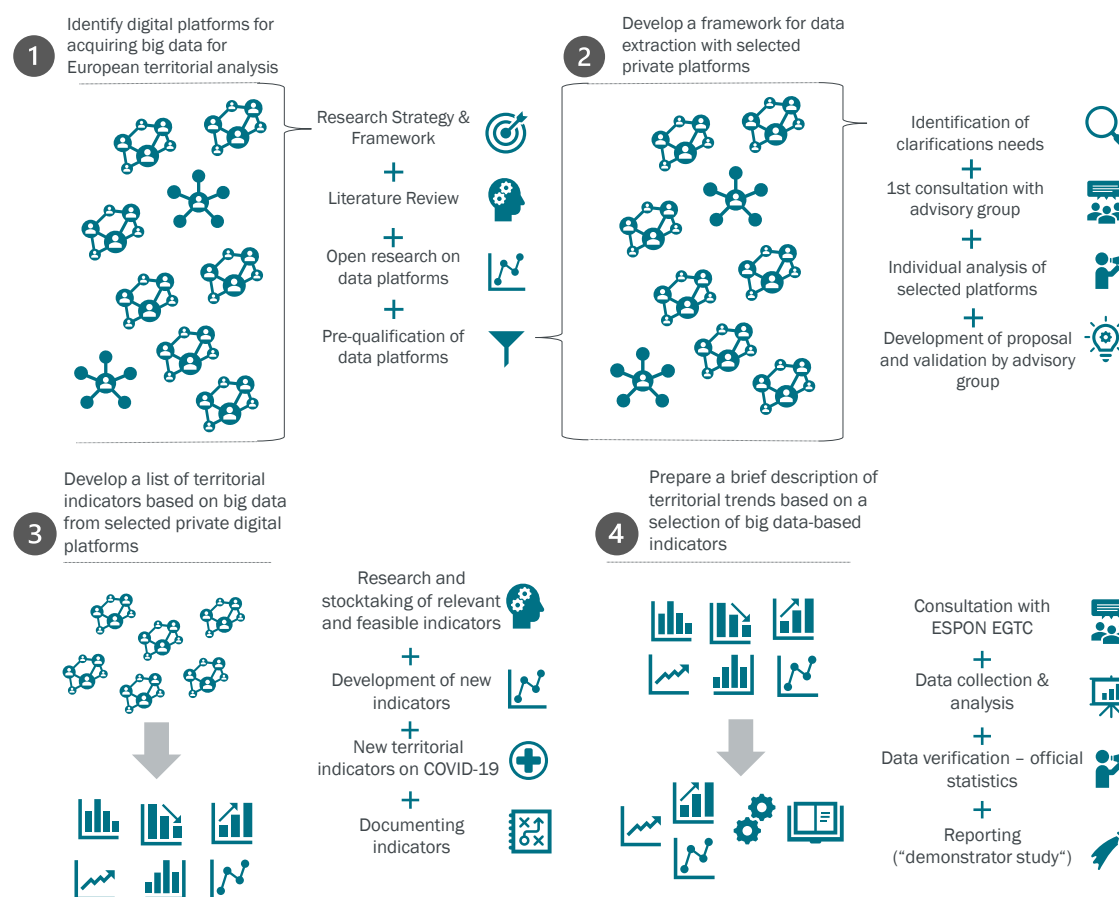
⁸ For instance, the Estonian central bank uses mobile phone data from Estonian mobile network operators to quantify inbound and outbound travel, and credit card payment data to calibrate expenditure figures.

⁹https://www.territorialagenda.eu/files/agenda_theme/agenda_data/Territorial%20Agenda%20documents/TerritorialAgenda2030_201201.pdf

1.2 Overall approach of the study

This study follows an explorative and experimental philosophy to respond to the core objectives described above. While some tasks were rather desk research driven, others included expert discussion and the use of more sophisticated analytical tools (web crawling, text mining, etc.). The following figure provides an overview of the **project's overall approach**, which centres around four key steps.

Figure 1-2: Overview of the project approach



Source: Prognos AG/DevStat (2021).

It should be highlighted that two focus groups with an advisory board (consisting of experts from the UN, Eurostat, academic experts, national statistical authorities, and others were held on the 27 May and the 1 July 2021), that contributed significantly to the work. The outcomes of this are further illustrated in the Annex and informed the content of the following chapters. Furthermore, valuable inputs were received from Prof. Dr. Heckmann (TU München) on the legal aspects around accessing data of private digital platforms (see also Chapter 3).

1.3 Structure of the report

The Final Report provides a final list of private big data platforms that are considered relevant (Chapter 2). Chapter 3 provides an analysis of the legal aspects of accessing data, an analysis of the terms of use of the selected platforms, and a structured framework of cooperation with insights on how to initiate cooperation. Chapter 4 presents a list of territorial indicators that could be developed using big data from the selected platforms and information on how to interpret the data. A brief description of territorial trends based on selected indicators is given in Chapter 5. Chapter 6 concludes with the main findings of the study.

Table 1-1 summarises the content for each chapter.

Table 1-1: Structure of the Final Report

Section	Title	Description
Chapter 1	Introduction	Providing an overview of the background and scope of the study as well as illustrating the objectives and related outcomes.
Chapter 2	Short list of private digital platforms	Overview of quantitative and qualitative analysis of platforms to come to a short list.
Chapter 3	Framework: Analysis of the legal aspects of access to data of selected platforms	Detailed analysis of legal aspects of different data access options and possible challenges of the five selected platforms.
Chapter 4	Territorial indicators based on big data	Presentation of an approach for developing a list of territorial indicators
Chapter 5	Demonstrator study	Illustration of how big data can be utilised for territorial analysis in practice

Source: Prognos AG/DevStat (2021).

2 Overview of relevant private digital platforms for territorial analysis

Chapter 2 has the key objective of presenting a list of private digital platforms. Moreover, it includes an in-depth description of how the relevant platforms were identified (quantitative and qualitative assessment). Thereby, the qualification process that was applied for the final selection of five platforms from a long list of private digital platforms will be outlined in detail.

2.1 Prequalified long-list of private digital platforms

This study started with a long list of 79 private digital platforms (see the annex – Table 7-1). These were selected considering the definition of **digital platforms as intermediates that allow different stakeholders (partners, providers, consumers) to share, extend and enhance digital processes and capabilities using a common digital technology system.**¹⁰ Platforms such as these give access to big data from the following fields:¹¹

- Access to information/content such as general search engines (e.g., Google) or specialised search engines (e.g., TripAdvisor). This category also includes services that grant access to other content, for instance, maps or video platforms;
- Access to personal data and other content such as social networks (e.g., LinkedIn);
- Access to goods and/or services such as online markets (e.g., Amazon) or sharing economy platforms (e.g., Airbnb);
- Access to the workforce or expertise / intellectual capabilities;
- Access to money or capital such as crowdfunding sites (e.g., Kickstarter, Gofundme) or payment systems.

Through this selection procedure, these **79 private digital platforms** were identified as potentially relevant providers of big data for the cause of this study. After the initial identification and classification of the platforms, a process of prequalification was conducted to filter the most relevant platforms. This process was based on certain criteria:

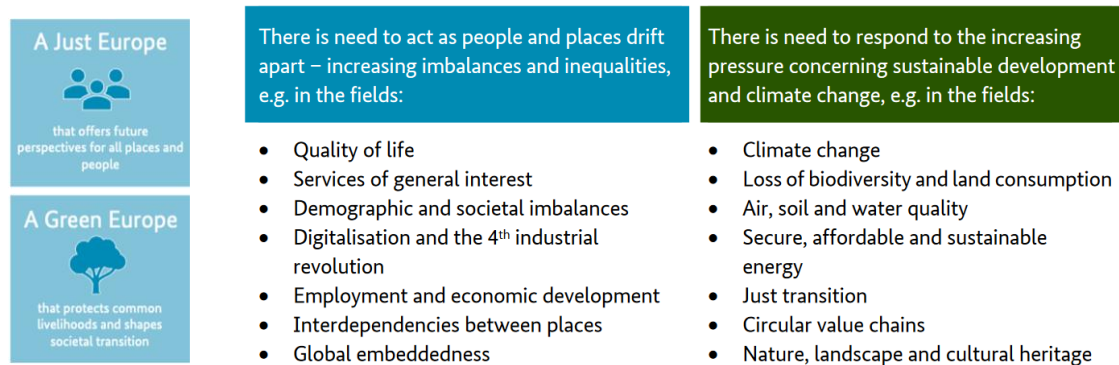
1. The platforms were sorted by their **availability of territorial traces / territorial data traces**. In the further process platforms that can only provide indirect territorial traces, were no longer taken into consideration for the prequalified longlist. The video-sharing platform YouTube, for instance, was not selected in this step since it does not allow for the extraction of territorial data traces without access to the user's Internet Protocol Address.
2. Since the development of new meaningful territorial indicators is a key outcome of this project, all remaining platforms that are not considered as **thematically relevant for ESPON** purposes were not selected for the prequalified long list. Examples of platforms that were not further considered due to their lack of thematic relevance are the streaming services Spotify and Netflix. For this purpose, the fields outlined in the Territorial Agenda 2030¹² (see Figure 2-1) were applied/connected to the private digital platforms.

¹⁰ European Commission (2016): Digital Platform for public services. Final Report. <https://joinup.ec.europa.eu/sites/default/files/document/2018-10/330043300REPJRCDigitalPlatformsBM-D2.5FinalReportv051018.pdf>

¹¹ Strowel, Alain; Vergote, Wouter (2016): Digital Platforms: To Regulate or Not To Regulate? https://ec.europa.eu/information_society/newsroom/image/document/2016-7/ucloouvain_et_universit_saint_louis_14044.pdf

¹² https://www.territorialagenda.eu/files/agenda_theme/agenda_data/Territorial%20Agenda%20documents/TerritorialAgenda2030_201201.pdf

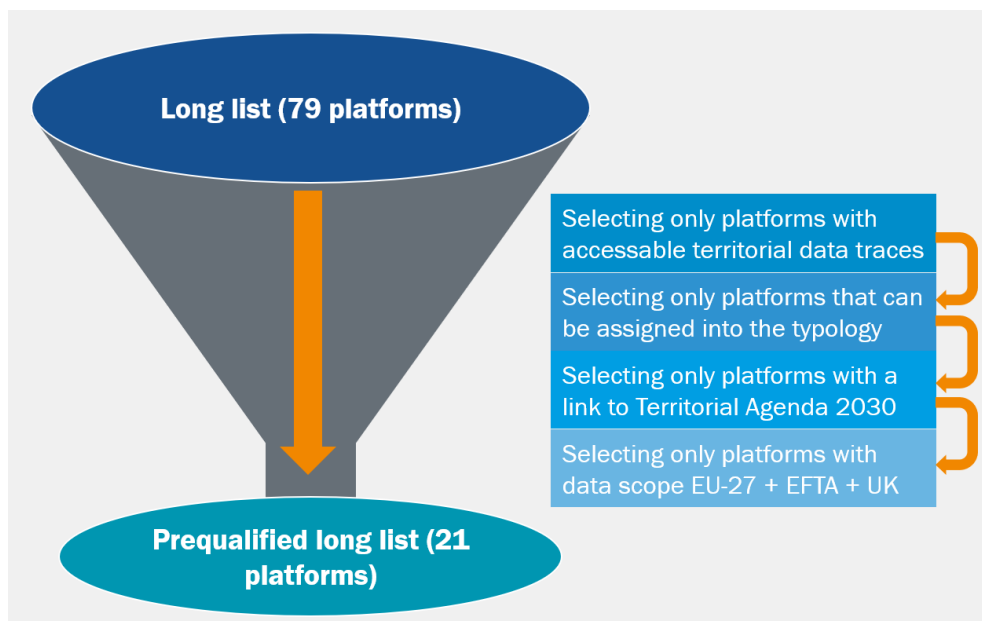
Figure 2-1: Territorial Agenda 2030 – a reference point for this project



Source: Prognos AG/DevStat (2021).

3. The remaining platforms were assessed regarding their **typology**. This was done to check if the platform constitutes a relevant big data platform that enables value-creating interactions between participants. In this step, platforms, that could not be assigned into any of the subcategories (e.g., access to personal data and other private content) were not selected for the prequalified long list. For instance, companies like the courier DHL or TomTom are developers of traffic and navigation software. Hence, they are potential providers of relevant territorial big data. However, these companies could not be assigned into any of these subcategories and were therefore not selected for the prequalified long list. It needs to be highlighted, that the platforms with no assignment to the typology, are not completely left out for further considerations since these platforms could potentially supply big data from which meaningful territorial indicators.
4. A key objective of this study is to identify digital platforms suitable for acquiring big data for **European territorial analysis**. Hence, in the last stage of the qualification process, only the remaining platforms whose data scope covers all 27 EU Member States, UK, and countries under the EFTA were considered for the prequalified longlist. For instance, the platform eBay was not selected in this step since it is only available in ten Member States. This process is further visualised in Figure 2-2.

Figure 2-2: Schematic overview over the qualification process



Source: Prognos AG/DevStat (2021). Own illustration.

The outcome of this process was a prequalified long list of **21 private digital platforms** that fulfil the outlined criteria. As Microsoft Academic will be discontinued by 2022, the platform will not be considered in the following steps, thus Figure 2-3 shows the remaining **20 preselected platforms**. The prequalified long list was presented and validated in the context of the **launch event of the advisory group** that was held on April 30th, 2021.¹³

Figure 2-3: 20 pre-selected private platforms

20 Platforms in the prequalified long list		
Knowledge platforms	News/Information/ Picture Sharing	Tourism platform
Google Scholar	Facebook Tumblr Twitter	Expedia Foursquare Trivago TripAdvisor Booking
Knowledge and Social media platforms	Route planning/ location identification	Picture Sharing
Academia ResearchGate	Google Maps	Flickr Instagram
Finance	Sports	Employment
PayPal	Strava	LinkedIn
Mobility	Lodging	
Kayak skyscanner	Airbnb	

Source: Prognos AG/DevStat (2021). Own illustration.

In line with the outlined criteria, all platforms contain certain touch points with the topics outlined in the Territorial Agenda 2030. These touch points are analysed in more detail in the following Section 2.2, which will be the basis to come to a short list of private digital platforms that will be used for data extraction.

2.2 Final list of private digital platforms

Building upon the prequalified long list of private digital platforms, this step aims to come to a short list of selected private digital platforms that are then used for developing a list of territorial indicators based on big data. Therefore, a quantitative, as well as a qualitative examination of the 20 platforms, was conducted. The **quantitative analysis** examines the platforms with the regard to their linkage to the fields of the Territorial Agenda 2030. The **qualitative part** analyses the platforms regarding user groups, data access possibility, and a selection of the type of analysis for the extraction of indicators from the data.

2.2.1 Criteria for the short list

For the short list, it is important to examine a broad spectrum of platform types to have different kinds of raw data sources for the analysis and developing indicators. For this purpose, platforms were classified by type so that platforms with similar functions are analysed together. The platforms are classified into the following types:

¹³ Points of discussion were among others the criteria used to select and qualify private digital platforms and first used cases. The next steps (Focus Groups 1 and 2) were outlined. The content of Focus Group 1 will be picked up in more detail in Chapters 3 and 4 of this report.

- micro blogging platforms,
- social media platforms,
- platforms regarding tourism,
- knowledge platforms as well as
- a platform regarding finance.

In the following, each type of platform is analysed qualitatively as well as quantitatively to be able to make a reasoned decision on which platforms to include in the short list.

Table 2-1 shows the overview of the **qualitative analysis** of each platform. The first column contains the primary user groups of the platforms. The second column gives insights into the kind of information one could get from this platform. Based on these two columns the third column gives ideas about what can be analysed. The table is based on expert evaluations, which were supplemented and expanded by desk research. Complementary, a **quantitative analysis** was performed to provide an orientation to which extent thematic fields of the Territorial Agenda are covered by private digital platforms. To that end, the occurrence of the thematic field on each platform has been examined by the means of an ontology. The ontology is oriented towards the Territorial Agenda and was developed manually based on the thematic fields. For example, in the thematic field of "digitalization and the 4th industrial revolution", terms such as "digitalization", "data protection", or "digital transformation" were requested.

To perform this analysis, the *prognos web intelligence tool* was used. The ontology was matched with the URL of each platform. With 20 platforms and 14 thematic fields the matching approach results in 280 combination rows as input for the *prognos web intelligence tool*. The *prognos web intelligence tool* approach is based on the Google API, which offers insights into hits of google search queries. In this case, the result indicates to which extent the terms of a thematic field are represented on a platform compared to another platform. This means that Google estimates how often a term such as digitalisation occurs on a platform and returns this value as a result. The results were aggregated by platform and thematic field so that there is one value per platform and thematic field. The platforms were then sorted by type as described above, for example, Twitter and Tumblr as micro blogging platforms. As the results are estimated values from Google and therefore partly very high and unmanageable, they have been transferred into clearer and more comparable values. For this purpose, the values were divided into quantiles for each of the types of platforms. This means that the quantiles only refer to the values of the platforms that belong to a type (see Table 2-2, Table 2-3, Table 2-4, Table 2-5). Quantiles represent a value below which a certain percentage of the values in its frequency distribution fall. For example, the 50th quantile is the value below which 50% of the values in the distribution are found. Values above the 90% quantile value were assigned to value 4. Values between the 90% quantile value and the 66% quantile value were assigned the value 3. Values between the 66% quantile value and the 33% quantile value were assigned to value 2. Values below the 33% quantile value and greater than 0 were assigned to value 1. The value 0 remains 0. This classification gives a good overview of which correlations are particularly high (value 4) and presents trends of correlation across all platforms by type. The values can be understood as comparative values. Since not all interesting data (e.g., financial transactions, routes, etc.) can be found directly on the website as terms, the quantitative approach also has its limitation here.

2.2.2 Qualitative and quantitative analysis by type of platform

Overall, the results of the quantitative analysis show, that the thematic fields "quality of life", "Nature, landscape, and cultural heritage" as well as the fields around "climate and environmental protection" have high values across all platforms. This may be since environmental and social sustainability are much-discussed topics and these discussions are also strongly reflected on the web. Moreover, the thematic fields "Services of general interest", "Interdependencies between places" and "Global embeddedness" tend to have lower values across all platforms. It can be assumed that these topics are less directly discussed than other topics on the platforms. Therefore, a qualitative analysis of these topics is more suitable.

Table 2-1: Platform profiles by type of platform

Type of platform	Platforms	Primary user groups	Insights into...	Analysis of...
A: Micro blogging	Twitter	<ul style="list-style-type: none"> journalists politicians publicists news consumers event participants and organizers 	<ul style="list-style-type: none"> political events reports on scientific results self-representation of stakeholders socially relevant issues and news communication of users 	<ul style="list-style-type: none"> news (nowcasting) (real-time) discussions to a wide range of topics and trends opinions and attitudes
	Tumblr	<ul style="list-style-type: none"> younger users private individuals who express their interests in form of blogs & pictures 	<ul style="list-style-type: none"> interests/hobbies of individuals self-representation of individuals 	<ul style="list-style-type: none"> pictures topics
B: Social Media	LinkedIn	<ul style="list-style-type: none"> employees/professionals employers (Senior-level) influencers human research manager marketing manager salespeople 	<ul style="list-style-type: none"> news and developments in firms relevant topics of certain professional groups self-reporting of career steps 	<ul style="list-style-type: none"> job profiles professional networks demographics in occupational groups interests and hot topics of professionals
	Facebook	<ul style="list-style-type: none"> private individuals who express their interests and opinions in the form of pictures, events, texts, etc. businesses that want to market their products 	<ul style="list-style-type: none"> networks of individuals interests of individuals 	<ul style="list-style-type: none"> private networks pictures opinions and attitudes interests
	Flickr		<ul style="list-style-type: none"> networks of individuals presentation of pictures 	
	Instagram		<ul style="list-style-type: none"> self-representation of individuals 	
C: Maps	Google Maps	<ul style="list-style-type: none"> people with the need for navigation people looking for entities (restaurants, medical offices, railway stations, etc.) 	<ul style="list-style-type: none"> search queries by location routes ratings of entities 	<ul style="list-style-type: none"> entities paths flows
	Strava	<ul style="list-style-type: none"> sportspeople 	<ul style="list-style-type: none"> fitness data of individuals routes 	<ul style="list-style-type: none"> paths sports activity

Type of platform	Platforms	Primary user groups	Insights into...	Analysis of...
D: Tourism	Foursquare	<ul style="list-style-type: none"> • holidaymaker • (non-business) travellers • tourists • restaurant visitor • individuals planning leisure activities 	<ul style="list-style-type: none"> • ratings of entities like hotels • supply and demand of flies, hotels, restaurants, apartments, package holidays, etc. 	<ul style="list-style-type: none"> • popularity places • tourist flows • possibilities of activity of places • distribution of entities like restaurants
	TripAdvisor			
	Booking			
	Airbnb			
	Trivago			
	Expedia			
	Kayak			
	Skyscanner			
E: Knowledge platforms	Research-Gate	<ul style="list-style-type: none"> • researchers • scientists • analysts 	<ul style="list-style-type: none"> • scientific topics • research institutes • researchers and scientists 	<ul style="list-style-type: none"> • scientific thematic focuses at specific locations and educational institutions • cross-site collaboration of scientists
	Academia			
	Google Scholar			
F: Finance	Paypal	<ul style="list-style-type: none"> • buyer • seller 	<ul style="list-style-type: none"> • money transactions 	<ul style="list-style-type: none"> • cash flows

Source: Prognos AG/DevStat (2021).

Below a description of the core findings by **type of platform** is presented:

A: Micro blogging platforms

The micro blogging platforms can be used to analyse **news, discussions to a wide range of topics and opinions, and attitudes** and therefore offer the advantage to analyse dynamic and rapidly evolving topics as well as multi-perspective topics. Micro blogging offers a variety of analysis options. Text can be analysed by statistical analysis methods but also hashtags and links could be interesting e.g., for a hyperlink analysis. Twitter has a stronger focus on real-time communication whereas Tumblr focuses on blogging over "replying" to other peoples' posts.

A comparison of the two platforms Twitter and Tumblr shows that Twitter has higher values overall. This could be because Twitter has a higher number of users. In addition, networks like Tumblr use tweets to distribute their content.¹⁴ By analysing Twitter it is possible to tap into topics of other platforms, for example by evaluating tweets or hashtags. Therefore, **Twitter** is selected as a micro blogging platform for the short list.

Table 2-2: Quantitative analysis of micro blogging platforms

A Just Europe							
Micro blogging platform\thematic field	Quality of Life	Services of general interest	Demographic and societal imbalances	Digitalisation and the 4th industrial revolution	Employment and economic development	Interdependencies between places	Global embeddedness
twitter.com	4	1	2	3	3	1	2
tumblr.com	3	1	1	1	2	1	1

A Green Europe							
Micro blogging platform\thematic field	Air soil water	Climate change	Loss of biodiversity and land consumption	Secure, affordable and sustainable energy	Just transition	Circular value chains	Nature, landscape and cultural heritage
twitter.com	3	4	3	3	2	2	2
tumblr.com	2	3	2	2	1	2	1

Source: Prognos AG/DevStat (2021).

B: Social media platforms

Social media platforms are websites and applications that enable users to create and share content or to participate in social networking which is why **network analysis and further statistical analysis regarding networks** are particularly suitable as analysis methods. It stands to reason that by analysing for instance opinions, attitudes, interests, and feelings can gain insight into multi-perspective topics. This creates the opportunity to introduce supplementary forms of data and measurement in addition to the established measures of the classic statistical offices, especially where data is missing or very difficult to obtain.

Looking at the individual platforms, the intended uses differ somewhat. While LinkedIn is a social network, which specialises in professions and careers and aims at networking highly qualified professionals, Instagram, Facebook, and Flickr focus more on networking private individuals. The data and insights that can be drawn from these social networks differ accordingly. Instagram, Facebook, and Flickr allow more of an analysis of interests and opinions such as preferences regarding vacation spots, events, food, hobbies. From LinkedIn, analyses can be derived on the topics of employment, qualifications, job profiles in demand, e.g., in the area of digitalization, etc (qualitative analysis). This is also reflected in the quantitative analysis. Here, LinkedIn shows a high correlation with "Digitalization and the 4th industrial revolution" as well as "Employment and economic development".

For this reason, **LinkedIn** is included in the short list as it can give highly interesting insights into the labour market, job profiles, highly relevant scientific topics, and demographics in occupational groups.

¹⁴ Severo, M. et al. (2015): Twitter data for urban policy making: an analysis on four European cities. https://www.researchgate.net/publication/279175120_Twitter_data_for_urban_policy_making_an_analysis_on_four_European_cities

Table 2-3: Quantitative analysis of social media platforms

A Just Europe							
Social Media platform\thematic field	Quality of Life	Services of general interest	Demographic and societal imbalances	Digitalisation and the 4th industrial revolution	Employment and economic development	Interdependencies between places	Global embeddedness
linkedin.com	4	2	3	4	4	1	2
facebook.com	4	3	3	3	3	2	2
flickr.com	2	1	1	2	2	1	1
instagram.com	2	1	1	1	2	1	1
foursquare.com	2	1	1	1	1	1	1

A Green Europe							
Social Media platform\thematic field	Air soil water	Climate change	Loss of biodiversity and land consumption	Secure, affordable and sustainable energy	Just transition	Circular value chains	Nature, landscape and cultural heritage
linkedin.com	3	3	3	4	2	3	3
facebook.com	3	4	3	3	3	4	3
flickr.com	2	3	3	2	1	2	2
instagram.com	2	2	2	2	1	2	2
foursquare.com	1	2	1	1	1	2	1

Source: Prognos AG/DevStat (2021).

C: Maps platforms

The maps platforms represent a special role due to their explicit territorial reference. They can give an insight into **the location and thus the distribution of entities** but also into the use of routes or flows of people. For the maps platforms, the table of quantitative analysis is omitted, since objects of investigation like routes or flows are not shown in it and individual terms on which the platforms are examined have little meaning for maps platforms. Since Google Maps is the most used online map service and Strava has a comparatively small number of users, we include **Google Maps** in the short list. Google Maps can also have a supporting function for territorial analysis (e.g., in case of ambiguity, measurement of distances, etc.)

D: Tourism platforms

Like the social media platforms, the tourist platforms provide advantages when it comes to analysing for instance **opinions, attitudes, interests**. Here, as well, the data can be used to supplement and expand the measures of traditional statistical offices that are not available or can only be obtained at high cost and effort. Unlike social media, it is less about networking and more about planning, booking, and evaluating leisure activities such as going on vacation, visiting restaurants, hotels, etc.

Table 2-4: Quantitative analysis of platforms regarding tourism

A Just Europe							
Tourism platform\thematic field	Quality of Life	Services of general interest	Demographic and societal imbalances	Digitalisation and the 4th industrial revolution	Employment and economic development	Interdependencies between places	Global embeddedness
tripadvisor.com	4	2	2	2	3	2	2
airbnb.com	4	2	2	2	3	1	1
booking.com	3	1	2	2	3	2	2
kayak.com	3	0	1	2	2	1	0
trivago.com	3	1	1	2	2	0	1
skyscanner.com	3	0	2	1	2	0	1
expediagroup.com	2	0	0	1	1	0	0

A Green Europe							
Tourism platform\thematic field	Air soil water	Climate change	Loss of biodiversity and land consumption	Secure, affordable and sustainable energy	Just transition	Circular value chains	Nature, landscape and cultural heritage
tripadvisor.com	4	4	4	3	2	4	3
airbnb.com	3	4	3	3	1	4	3
booking.com	3	3	3	3	1	4	3
kayak.com	3	3	2	2	1	3	2
trivago.com	2	2	2	2	2	3	2
skyscanner.com	2	2	3	2	0	3	2
expediagroup.com	1	1	1	1	1	1	1

Source: Prognos AG/DevStat (2021).

A quantitative comparison of the platforms shows that TripAdvisor, Airbnb, and Booking have higher hit rates than Trivago, Skyscanner, and Expedia. Since Airbnb makes access to its API difficult or does not allow commercial use of the data, it is excluded. TripAdvisor provides insights into many dimensions of leisure offers and activities. Therefore, we consider **TripAdvisor** the most suitable platform for our short list.

E: Knowledge platforms

The analysis of knowledge platforms can provide a detailed **insight into scientific topics and trends**. The advantage of a territorial analysis is the assignment of scientific work to institutes, so that scientific objects of study, scientific collaboration, and cooperation can be localized. The analysis of data from knowledge platforms can give insights into dynamic and rapidly evolving scientific issues.

Overall, ResearchGate has the comparatively highest hit rate in qualitative analysis. We, therefore, assume that **ResearchGate** offers the most comprehensive data and was selected for the short list.

Table 2-5: Quantitative analysis of knowledge platforms

A Just Europe							
Knowledge platforms platform\thematic field	Quality of Life	Services of general interest	Demographic and societal imbalances	Digitalisation and the 4th industrial revolution	Employment and economic development	Interdependencies between places	Global embeddedness
researchgate.net	4	3	3	3	3	2	2
academia.edu	3	2	2	2	3	2	2
scholar.google.com	2	1	2	2	2	1	1

A Green Europe							
Knowledge platforms platform\thematic field	Air soil water	Climate change	Loss of biodiversity and land consumption	Secure, affordable and sustainable energy	Just transition	Circular value chains	Nature, landscape and cultural heritage
researchgate.net	3	4	4	4	3	3	3
academia.edu	3	3	3	3	2	3	3
scholar.google.com	2	3	2	2	1	2	2

Source: Prognos AG/DevStat (2021).

F : Finance platform

The finance platform Paypal can offer insights into money transactions as cash flows can be analysed. Data from Paypal can give insight into a thematic field related to flows like “Interdependencies between places” and “Global Embeddedness”. However, it is difficult since individual terms do not play a role in such exploitation. Since data is very difficult to obtain due to data protection, we suggest **not including Paypal** in the short list.

2.2.3 Short List of selected platforms

Table 2-6 summarises the selected platforms by type as well as the key findings of the qualitative and quantitative analyses from the previous chapter. The data accessibility of the platforms is analysed in Chapter 3.

Table 2-6: Summary of selected platforms

Twitter	LinkedIn	Google Maps	TripAdvisor	ResearchGate
Micro blogging	Social media	Maps	Tourism	Knowledge platform
Qualitative Analysis: Objects of analysis ...				
<ul style="list-style-type: none"> News Discussions of a wide range of topics, opinions, attitudes 	<ul style="list-style-type: none"> Networks Labour markets Job profiles Highly relevant scientific topics 	<ul style="list-style-type: none"> Location distribution of entities Routes or flows of people 	<ul style="list-style-type: none"> Location Opinions, attitudes interests related to locations Leisure activities 	<ul style="list-style-type: none"> Scientific topics and trends Scientific collaboration/cooperation
Quantitative Analysis: Top 3 of thematic fields ...				
<ol style="list-style-type: none"> Quality of life Climate Change Biodiversity, Digitalisation & more 	<ol style="list-style-type: none"> Digitalisation & the 4th industrial revolution Quality of life Employment 	Quantitative analysis omitted → relevant information not available in form of words	<ol style="list-style-type: none"> Circular value chains Quality of life Climate change 	<ol style="list-style-type: none"> Quality of life Climate change, biodiversity & sustainable energy Cultural heritage

Source: Prognos AG/DevStat (2021).

3 Framework: Analysis of the legal aspects of access to data of selected platforms

This chapter is dedicated to a **detailed analysis of legal aspects of access to these selected platforms**. Moreover, different data access options and possible challenges of the five selected platforms will be outlined in-depth. The focus group with the advisory board that was held on the 27th of May with the topic 'Difficulties and specificities regarding data access and agreements and communication with platforms' significantly contributed to this chapter, as topics such as challenges and difficulties in accessing big data and establishing cooperation agreements were discussed. In addition, the legal aspects of access and processing of the data were critically reflected by the project team's senior legal expert Professor Dr. Heckmann.

3.1 General information on accessing data of platforms

For the legal analysis of data access and data processing with regards to large internet platforms, it is important to first look at the target **object of access**. Very different legal questions are raised depending on whether, for instance, it is a matter of claims for information against the platform operator, the processing of individual pieces of publicly accessible information, or **full access to the data aggregated there within the platform's functionality** (in the sense of a "big data analysis").

The present study is aimed solely at the latter variant and has a focus on specific use cases. Even if the specific data sets sought may still be individualized by a defined query mode or using filter technologies (depending on the platform) it is noticeable that, at least for the present study and its contexts of use, there is neither a clear exclusion of personal data nor a clear exclusion of sensitive data requiring special protection.

3.1.1 Entitlement to data access

The first legal question to be clarified concerning data access to selected large Internet platforms is that of a right to data access based on European or national legal norms. Should such a right already exist (*de lege lata*) or should it be adopted soon (*de lege ferenda*), the strategy of data access would have to be aligned with the legal basis and, in particular, fulfil the requirements that may be regulated therein.

Currently, there is no such general data access claim at present, nor are there any apparent political plans to introduce it soon. There **is currently no explicit statutory right to** access the data of private platform operators.¹⁵

There has been repeated talk of a **data-sharing obligation**, for example in the political discussion in Germany that accompanied the preparation of the Federal Government's 2020 data strategy. However, the way in which this data strategy was finally formulated in the Federal Chancellery and adopted in the Federal Cabinet on 27 January 2021 Germany relies on a **system of incentives for data sharing** and rejects a legal obligation for companies, especially platform operators.¹⁶

This is no different at the European level. Here, the **Data Governance Act** is also to be understood as a rather voluntary concept for data access. Its purpose is to **improve data sharing between companies**, to enable the use of personal data with the help of a data intermediary, and to support the altruistic use of data

¹⁵ see answer of the Federal Government to the 4th question in the small inquiry on data access to social media platforms for research purposes, BT-Drs. 19/10595, 03.06.2019, p. 4; available at: <https://dsrserver.bundes-tag.de/btd/19/105/1910595.pdf>.

¹⁶ "We are promoting a culture of willing and responsible data sharing for the benefit of everyone in society." Federal Government of Germany (2021): Datenstrategie der Bundesregierung, p.22. Available online: <https://www.bundesregierung.de/resource/blob/992814/1845634/f073096a398e59573c7526feaaadd43c4/datenstrategie-der-bundesregierung-download-bpa-data.pdf> (in German; accessed on 08.09.2021)

("data donation").¹⁷ In the view of the EU Commission, data access could be regulated at the European level in certain areas if the Commission determines that competition in downstream markets cannot otherwise be ensured due to a lack of data access.¹⁸

There is a set of rules that refer to specific categories of data and their use in specific contexts. These include, for example, the **EU Commission's Guidelines for Strengthening the Code of Conduct on Combating Disinformation of 26.05.2021**:¹⁹ The Code of Conduct provides a voluntary obligation on the part of platforms to combat disinformation and, associated with this, the creation of transparency. To this end, it may be necessary to grant certain organisations, particularly those in the research community, access to data from the platforms so the data can be analysed and recommendations for action can be developed. Data scientists with relevant expertise are thus able to analyse datasets necessary for understanding sources, vectors, vectors, and dissemination patterns that characterise the disinformation phenomenon.

The **criteria according to which this data access is organised** are also interesting for other contexts of the use of data from platforms. Point 8.1.3 of the guidelines, for example, talks about standardised conditions that are uniform across platforms, about the quality of researchers and minimum categories of data that are made available, about technical, and organisational security measures for data processing and the prevention of reassignment in the case of pseudonymized data.

However, the data points required for a systematic analysis of public posts (technical data, reach data, search result data, etc.) can only be viewed to a very limited extent.²⁰ In addition, **research analyses of social media platforms are fundamentally in an area of legal tension** since the vast majority of the data examined in the process contains legally protected personal information and data.²¹

To the extent that a right to data access is envisaged, such as by Article 31 of the planned **Digital Service Act**²², this relates to a specific constellation that is not relevant in the present context. It concerns the obligation of "very large online platforms"²³ to provide the digital services coordinator at the place of establishment or the Commission access to the data necessary for the monitoring and evaluation of compliance with this regulation upon a reasoned request within a reasonable period specified therein. That digital services coordinator and the Commission shall use those data exclusively for those purposes. The same applies to access to data by researchers for the sole purpose of conducting research that contributes to the identification and understanding of systemic risks. Access to data is provided by very large online platforms through online databases or application programming interfaces.

Art. 26 DSA considers as **systemic risks** in this sense the dissemination of illegal content through its services, possible adverse effects on the exercise of fundamental rights, and the intentional manipulation of the service, including through inauthentic use or automated exploitation of the service, with actual or foreseeable adverse effects on the protection of public health, on minors and on social debate, or actual or foreseeable effects on electoral processes and public security.

3.1.2 "Contract resolution": access only to voluntarily provided data sets

Access to the data required for the project must be **provided voluntarily by the platform operators**. This occurs in a combination of technical access and contractually regulated access and exploitation conditions. In principle, this corresponds to the EU-wide **principle of private autonomy**, according to which it is up to

¹⁷ Hartl and Ludin (2021): "Recht der Datenzugänge", MMR, 534 (537).

¹⁸ Hartl and Ludin (2021): "Recht der Datenzugänge", MMR, 534 (537).

¹⁹ Available at: <https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:52021DC0262&from=de>. (last accessed 08.09.2021)

²⁰ see. www.stiftung-nv.de/sites/default/files/blinde.fleck_.digitale.oeffentlichkeit.pdf.

²¹ Preliminary note by the questioners of the small question on data access to social media platforms for research purposes, BTag-Drs. 19/10595, 03.06.2019, available at: <https://dserver.bundestag.de/btd/19/105/1910595.pdf>.

²² see <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package>. (last access 08.09.2021)

²³ These are platforms "that, because of their reach, play a central, systemic role in fostering public debate and economic transactions." see <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package>. (last access 08.09.2021)

private companies and their business partners and customers to determine the conditions of their economic cooperation themselves and to act accordingly.

Such private autonomous action is often also characterised by pragmatism, balancing of interests, and solution orientation. For example, the platform operators guarantee technical access by creating an interface (API).

Box 3-1: Application Programming Interface - definition

i

Application Programming Interface (API)

An API is an interface that is offered by many websites or apps. This gives easy-to-use access to certain resources of the database behind the website or app. Thereby, an API has an intermediary function and makes the websites or apps data accessible to third parties for testing, data analytics, and product development. In other words, an API allows two different applications to communicate with each other. Usually, the platforms also communicate their terms of use in connection with the provision of an API. Since this very important information is publicly available, it constitutes an **excellent basis for developing frameworks of cooperation**.

Frequent restrictions can be **API rate limits** that restrict the number of possible data requests in a given amount of time. Attention was also paid to the respective **pricing models** for the data access and regulations regarding the accessibility of the APIs. In this context, the information provided by the platforms in their web pages dedicated to developers and the respective terms of use is a crucial starting point for the analysis of technical aspects and other challenges.

Source: Prognos AG/DevStat (2021).

Other options would be to tolerate data access via so-called **web crawlers**. Restrictions and details of data access and data utilisation are regularly regulated in the **terms of use**, the approval of which leads to the conclusion of a **user contract**, which in turn forms the legal basis for further data handling. Similarly, data access via web crawlers can be linked to explicit instructions in the "robot.txt" file.

A complete regulation cannot be expected with this approach. In some cases, platform operators are also open to special agreements. Ultimately, however, it is precisely these pre-formulated and, if necessary, individually added **terms of use that form the legal framework for general data access**.

3.1.3 Restrictions on data access and data use

However, such a contractual data access and data use agreement may not be concluded at the expense of third parties. Therefore, property rights must be observed here, which relate in particular to the rights of other parties involved.

Data protection law:

This primarily concerns data protection law. Insofar as the requested data is **personal data or at least data that can be related to a person** (i.e., data where the identification of the persons affected is possible), the strict requirements of the General Data Protection Regulation (GDPR) apply. This means first and foremost that there must be a **justification for the data transfer and the further processing steps**. In particular, the following articles can be considered:

- the consent of the affected persons (Art. 6 para. 1 lit. a DSGVO)
- the balancing of interests (Art. 6 para. 1 lit. f DSGVO).

An individual consent of each affected person, especially after transparent information according to Art. 12 et seq. GDPR, is illusory. Only **approval by the platform operator** could be considered. The latter is not an affected person in the strict sense. However, it is conceivable that he granted himself the rights to the data in the context of the platform used by the affected persons (i.e., those who, for example, sent a tweet,

wrote a hotel review, or uploaded a document). Then, it would - at least at first glance - no longer depend on the consent of the person affected.

In fact, the platform operators selected in this study also largely **cede rights from their users**. Whether this is compatible with European law is problematic. Indeed, data is now also considered a "currency" under European law, so that a certain "commercialisation" of data is taking place. But whether this can be exempt from all the restrictions of the General Data Protection Regulation is questionable. Ultimately, this amounts to all users, as affected persons of any further processing of their personal data by third parties, waiving basic data protection rights such as the revocability of consent, rights of access and objection, etc. This is difficult to justify. It should also be borne in mind here that ultimately, high-level fundamental rights from Articles 7 (Respect for private and family life) and 8 (Protection of personal data) of the EU Charter of Fundamental Rights are at stake.

Against this background, one should rather try to base the processing of personal data on the **balancing of interests according to Art. 6 (1) lit. f DSGVO**. According to this, the data access and the related data processing are justified if it is "necessary for the purposes of the legitimate interests of the responsible person or a third party", "unless such interests are overridden by the interests or fundamental rights and freedoms of the affected person which require the protection of personal data". The data access in the present context is legitimate because there are reasonable interests for the analysis of these data files as set out in detail in the study (see Chapter 1) and do hence not need to be repeated here. Whether the interests of the affected persons prevail depends crucially on how the body accessing and further using these data handles personal information. Certainly, the body accessing the data cannot and will not simply ignore the fact that there is also personal data in the data pool. If, on the other hand, the body accessing the data designs the entire **data management in** such a way that the interests of the affected persons are taken into account by means of anonymisation and pseudonymisation measures, by replacing them with synthetic data or using filtering technologies that meet the interests of the affected persons, then these may be subordinate. Once again, data protection can be guaranteed through **technology design**.

It is generally assumed that the processing of publicly available (personal) data through **data scraping** is generally permitted on the basis of a legitimate interest (Art. 6 (1) (f) GDPR). However, the responsible person will generally have no choice but to **inform the affected people**. An exception to this notification obligation is only made under Art. 14 (5) (b) of the GDPR if the provision of the information proves impossible or would require a disproportionate effort. In practice, this is sometimes also handled strictly. For example, the Polish supervisory authority UODO imposed a fine on a Polish company that collected information on 6 million persons from public databases. The Polish regulator considered that there were cost-effective ways of informing even a large number of affected people about the data collection: for example, through short commercials before the main news, SMS messages or broad information on internet portals.

Further requirements, such as in particular the need for a **data protection impact assessment**, should only be pointed out here.

In any case, to the extent that data access can be limited to the **collection of purely factual data**, the General Data Protection Regulation is not applicable.

Copyright:

In addition to data protection law, copyright law is also relevant for data access and data exploitation in the present context.

The Copyright Directive (EU) 2019/790 (so-called **DSM Directive**), which came into force on 6.6.2019, contains, in addition to the much-discussed regulations on the ancillary copyright of press publishers or on platform liability, new regulations on **"text and data mining"** in Art. 3 and Art. 4 of the DSM Directive (DSM Directive).

According to Art. 4 DSM, the member states must provide for exceptions or limitations "for reproductions and extractions of lawfully accessible works made for text and data mining". Reproductions and extractions may be kept for as long as necessary for text and data mining. The exceptions and limitations apply "unless the respective rights holders have expressly reserved use of the [...] works [...] in an appropriate manner, such as by machine-readable means in the case of content published online."

In Germany, the Directive has been implemented, inter alia, in **Sections 44b and 60d UrhG**. According to Section 44b (1) UrhG, "text and data mining ... is **the automated analysis of single or multiple digital or digitised works in order to extract information therefrom, in particular about patterns, trends, and**

correlations.“ According to Section 44b (2) UrhG, text and data mining is only permissible if the reproductions relate to **lawfully accessible works**. The reproductions must be deleted when they are no longer required for text and data mining.

All uses in the sense of text and data mining are otherwise only permissible if the rightsholder has not reserved them. A **reservation of use** in the case of works accessible online is only effective if it is made in machine-readable form (Section 44b (3) UrhG). This is usually done utilizing the platforms' terms of use available online.

Furthermore, Section 60d UrhG regulates text and data mining for the **purposes of scientific research**. According to this, research organisations such as universities, research institutes, or other institutions conducting scientific research are entitled to reproduce for text and data mining purposes if they

- pursue non-commercial purposes,
- reinvest all profits in scientific research or
- operate in the public interest within the framework of a mission recognised by the State. This does not apply to research organisations that cooperate with a private undertaking that has a determining influence on the research organisation and privileged access to the results of scientific research.

For this aspect of data access in the sense of text and data mining, copyright law confirms the **practice of the platform operators** described above **when granting data access for** analysis purposes: the focus is on the terms of use specified by the platforms. If these permit data access and subsequent exploitation and reproduction, this is also acceptable under copyright law within the scope of this permission. Access outside of explicitly published terms of use can (in addition to the General Data Protection Regulation) also violate copyright law.

3.1.4 Key legal aspects for data access

The legal analysis has shown that there is no statutory right for accessing data of private big data platforms. Instead, **data access must base on agreements with the platform operators. These agreements can either base on individual agreements or on pre-formulated agreements via the platforms terms of use.** A critical aspect in terms of data use concerns personal data. Insofar as the obtained data sets in also contain personal data, it is questionable whether their processing is already justified by the fact that the users of the platforms have granted the operators far-reaching rights of use concerning this data. The extent to which stricter requirements are to be placed on the inevitably associated waiver of the assertion of rights of the persons affected under the GDPR has not yet been conclusively clarified in legal terms. It is therefore recommended that data processing be based on a balancing of interests pursuant to Article 6 (1) (f) of the GDPR and that the interests of the persons affected be sufficiently taken into account through an appropriate design of the data access. This means that data that is used for the construction of territorial indicators needs to be treated by, for instance, anonymization or pseudonymization measures if the data acquired contains personal data. Such Personal data is, for instance, encountered when accessing Twitter data through the Twitter API where data is partially connected to usernames that in some cases can be connected to persons or organisations. Data access is unproblematic as far as it concerns purely factual data. Text and data mining can also be designed in accordance with copyright law, taking into account the specifications of the platform operator. Ultimately, the decisive factor is what access a platform grants and how it is used sensibly for one's analysis purpose.

There are **several possibilities for accessing the data of a private digital platform**. Most of them are indirect ways and are not based on more complex cooperation agreements.²⁴ For example, web scraping is a technology to extract certain information from (the texts of) a website. The robot's exclusion protocol defines which contents may be scraped.

²⁴ As the focus groups and other studies show, cooperation agreements are not necessarily the ideal solution because of the cost, lengthiness and complexity of the process. Most of the points are linked to the difficulty to engage partners in the first place, as they have no real incentive to share data.

However, the most important direct access is an API. Unlike indirect access, an **API avoids several methodological challenges**. Extracting data from a website always involves a certain amount of error and fuzziness. Direct access to a platform's databases avoids these problems. Another important advantage is that it is clear which data the platforms want to share with third parties and which not. The access options via API are further described in the following per selected platform.

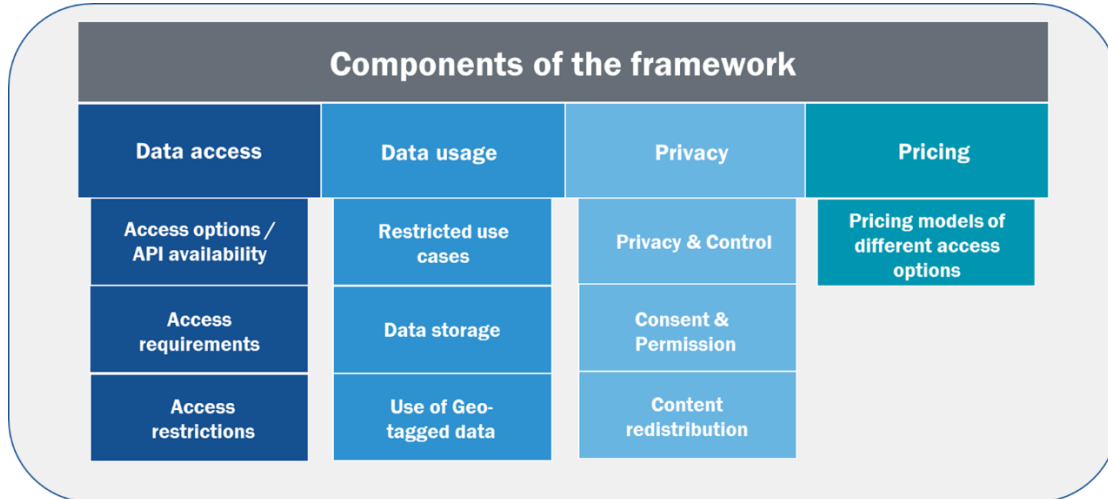
3.2 Analysing the specific requirements of data access and processing of the selected platforms

In this section, the specific requirements of the five selected platforms regarding their data access and data processing will be scrutinised in detail. The baseline for this analysis is the reflections on the legal aspects of data access that were presented in the previous section. To analyse the selected platform regarding their requirements of data access and processing a framework of cooperation was developed that structures the analysis. This framework will be presented hereafter followed by a specific examination of the selected platforms.

3.2.1 Components of the framework for for accessing data from private digital platforms

For the analysis of data access and data of the five selected platforms a framework of cooperation was developed that serves as a guideline for analysing platforms in this regard. An overview of the **components of this framework** is provided in Figure 3-1. This framework builds upon the reflection on the legal aspects of data access of the previous section as well as the specific requirements of this study that include, for instance, the use of geo-tagged data and the pricing models of the platforms. In additions, several components are listed under the dimensions of 'Data Access', 'Data usage' and 'Privacy'.

Figure 3-1: Component of the framework for cooperation for the analysis of data access & data usage



Source: Prognos AG/DevStat (2021).

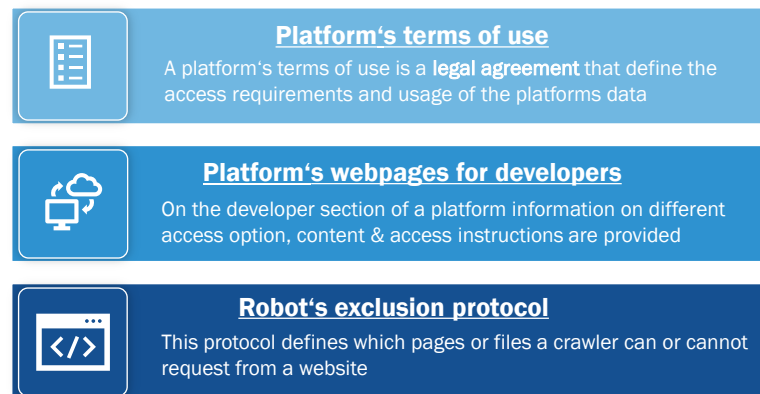
A feature of this framework for cooperation is that it cannot only be used to analyse the five previously selected platforms but also for analysing any other platform regarding its specific requirements for data access and data usage. In the following the different sources that are used to in the analysis of platforms regarding their data access and data usage are presented.

As it was described in the previous section the terms of use of platforms can be considered as a pre-formulated agreement with a platform that sets the terms under which data can be accessed and used. Therefore, an important source of information for the following analyses are the respective terms of use of the platforms. Since there is no entitlement to data access as outlined before it will first be checked whether the selected platforms do in fact permit access to their data. Moreover, since APIs constitute an excellent method (as described before) they will be the focal point of this section and therefore, the respective platform's

webpages for developers in which often the technical requirements of an API access are explained will be analysed. Where an access via an API is not possible an access using a web scraping approach is a viable approach. To analyse the specific requirements in such a case the robot's exclusion protocol of the platforms will be examined. This protocol or robots.txt file states a standard in the communication between websites and web crawlers. It clarifies which pages or files a crawler can or cannot request from a website.

The figure below provides an overview of the relevant and important information sources that are used to analyse the specific requirements of data access and processing of the selected platforms.

Figure 3-2: Relevant information sources



Source: Prognos AG/DevStat (2021).

The following sections illustrate the terms of use and the access options of the short-listed platforms Twitter (Section 3.2.1.), LinkedIn (Section 3.2.2.), Google Maps (Section 3.2.3), TripAdvisor (Section 3.2.4.), and ResearchGate (Section 3.2.5.).

3.2.2 Twitter

The microblogging platform Twitter presents an excellent source of territorial information through its geo-tagged content. In this section, the data access options, and possible limitations will be elaborated on. Twitter offers access to its data through **three different APIs²⁵ (Standard, Premium and Enterprise) that vary in terms of their content and their pricing** (see also Table 3-2).

Table 3-1: Overview of Twitter data access options

Terms of use permit external data access	API available	Other official data access interfaces/tools available	No significant access restrictions
✓	✓	✗	✓

Source: Prognos AG/DevStat (2021).

While the **Standard API** grants access to the tweets of the last seven days and geo-tagged information the **Premium API**, for instance, also allows accessing older and historic tweets. Furthermore, the different APIs also differ with regards to their rate limits and their pricing models. The Standard API is limited to 900 requests per 15 minutes and is free of cost. The Premium API allows for a higher rate limit but costs between \$99 and \$149 per month depending on chosen rate limit and content.

²⁵ <https://developer.twitter.com/en/docs>

Table 3-2: Factsheet Twitter API

Twitter API	
Available APIs	Standard API Premium API Enterprise API
API-Content	<p><u>Standard API:</u> Tweets (last 7-days), Users, Direct Messages, Geoinformation, etc.</p> <p><u>Premium API:</u> Access to historic Twitter data, more complex queries, higher rate limits, metadata enrichments</p> <p><u>Enterprise API:</u> Realtime and historical data, access to Engagement API (measuring and optimising content)</p>
Regional granularity	Depends (up to NUTS3 for geo-tagged tweets)
API rate limits	Guideline: 900 requests per 15-minutes 3200 tweets per account
Access	User registration and request for developer account required
Pricing	<p><u>Standard API:</u> free</p> <p><u>Premium and Enterprise API:</u> \$99 -149 per month (depending on content and rate limit)</p>

Source: Prognos AG/DevStat (2021).

The access to the Twitter APIs is well documented by Twitter.²⁶ In order to be granted access, it is **required to apply for a Twitter developer account**. In this application, the **specific use case** for the API must be presented. When applying for the Twitter API it is possible to choose between the **Standard and the Academic Research track**. The latter gives elevated access and enhanced functionality to academic researchers. However, a crucial requirement for a successful application to the Academic Research track is that the application is filed by a research-focused employee at an academic institution or university (see infobox). In this context members of the focus group expressed that in their experience the access to the Twitter API can be difficult depending on the needed data.

Box 3-2: Twitter Academic Research Track



Twitter Academic Research Track

With its Academic Research Track Twitter provides special data access to academia. This specialised track contains global, real-time, and historical data. In addition, this data access promises a higher monthly volume rate and enhanced features that are specifically designed to support research. Twitter provides three different access levels to the Academic Research Track (Basic, Elevated, and Custom) that differ with regards to their data volume, query rules, streaming rates, and costs. However, at this point the Elevated and Custom access levels are still under development, only the Basic access mode is currently available.²⁷

Requirements

The Twitter Academic Research Track is reserved for “Academic researchers with specific research objectives [...] This includes graduate students working on a thesis, Ph.D. candidates working on a dissertation, or research scholars affiliated with or employed by an academic institution.”²⁸ Applicants need to provide a clearly defined research objective and a specific plan for how the Twitter data will be used, analysed, and shared. Moreover, this type of data access is restricted to non-commercial use.

Source: Prognos AG/DevStat (2021).

In the **terms of use** of their APIs, Twitter sets out specific terms that have to be considered for the construction of territorial indicators using Twitter data. These will be presented in the following:

1. One requirement concerns the **handling of geographic information** which may not allow the storage and aggregation of location data. It is not allowed to separate location data from tweets: “*You may not separate location data or geographic information from Tweets to show where individuals have been over time. Heat maps and related tools that show aggregated geo activity (e.g.: the number of people in a city using a hashtag) are permitted.*”²⁹
2. Moreover, **offline stored data** needs to be kept updated with the content on Twitter. This means that, for instance, if geotags are removed from tweets this information also has to be removed in the data stored offline.
3. Another term by Twitter **forbids the use of the APIs to measure the usage of Twitter for benchmarking** or other competitive purposes: “*You may not use the Twitter API to measure the availability, performance, functionality, or usage of Twitter for benchmarking, competitive, or commercial*

²⁶ <https://developer.twitter.com/en/apply-for-access>; <https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api>

²⁷ As of September 2021.

²⁸ <https://developer.twitter.com/en/products/twitter-api/academic-research>

²⁹ <https://developer.twitter.com/en/developer-terms/agreement-and-policy>

purposes. For example, you should never use the Twitter API to: Calculate aggregate Twitter metrics, such as the total number of Monthly Actives (MAs) or Daily Actives (DAs), [...] Calculate aggregate Twitter Tweet metrics, such as the total number of Tweets posted per day, or the number of account engagements, etc."³⁰

Although the latter may not be the intention it still needs to be considered when constructing territorial indicators using Twitter data. For instance, an indicator that measures the total number of Tweets posted per day may be a violation of the Twitter API terms of use. Overall, the terms of use offer a promising option in accessing Twitter data via the platform's API. Nonetheless, **certain requirements (e.g., regarding the handling of geographic information) must be kept in mind when handling Twitter data and constructing indicators based on this data.**

3.2.3 LinkedIn

LinkedIn, a platform that is primarily used for professional networking, **provides a variety of different access options to their data.** Besides the access via API LinkedIn also provides access to specific tools, e.g. the LinkedIn Economic Graph³¹. These access options will be further discussed in the following.

Table 3-3: Overview of LinkedIn data access options

Terms of use permit external data access	API available	Other official data access interfaces/tools available	No significant access restrictions
✓	✓	✓	✓

Source: Prognos AG/DevStat (2021).

LinkedIn offers a variety of business solutions including Consumer-, Talent-, Marketing-, Sales- and Learning-Solutions, which also have an API. Several different products are united under each solution tool, e.g., the LinkedIn Talent Solution tool includes LinkedIn Recruiter, LinkedIn Jobs, and LinkedIn Talent, among others. The data that can be accessed through the LinkedIn APIs covers various employment-related information, for instance, on job postings, the mobility of employment, and employment by industries. LinkedIn provides historical (in case of the LinkedIn Economic Graph data dates to 2019³²) and locational data. The latter depends on the information provided by the users and the data provided in job postings. Like most platforms, LinkedIn specifies a maximum number of API calls that can be made in a certain period. The precise rate limit varies depending on the endpoint addressed by an API call and is reset every day.³³ Currently, all LinkedIn APIs are provided for free.³⁴

³⁰ <https://developer.twitter.com/en/developer-terms/agreement-and-policy>

³¹ <https://economicgraph.linkedin.com/>

³² <https://graph.linkedin.com/insights/labor-market>

³³ <https://docs.microsoft.com/en-us/linkedin/shared/api-guide/concepts/rate-limits>

³⁴ <https://legal.linkedin.com/api-terms-of-use>

Table 3-4: Factsheet LinkedIn API

LinkedIn API	
Available APIs	LinkedIn APIs (Consumer, Talent, Marketing, Sales and Learning)
API-Content	Depends on API (e.g., job postings, location, industry, etc.)
Regional granularity	Depends on user information (up to NUTS3)
API rate limits	API requests are rate-limited. Rate limits are reset every day and vary depending on which API endpoint is used
Access	User registration and request for developer account required. Depending on API more applications are required (e.g., Marketing API requires access to Marketing Developer Platform)
Pricing	free

Source: Prognos AG/DevStat (2021).

To access the LinkedIn data, it is mandatory to create a **developer account** and to have an authorized application:

“To access the APIs you must sign-in to LinkedIn, and register an Application. [...] Both the Application registration and your Member account must contain accurate and up-to-date information at all times, including your current title, company, and e-mail address. [...]. Once you have successfully registered an Application and met the other requirements for a particular API, you will be given the necessary credentials to access that API.”³⁵

Box 3-3: LinkedIn Economic Graph



LinkedIn Economic Graph

With its Economic Graph LinkedIn collaborates with governments and NGOs providing data on skills, jobs, and industries. So far, the LinkedIn Economic Graph has worked together with Eurostat and the World Bank (among others). Five key themes are in the focus of the LinkedIn Economic Graph: Sustainable Economy, Emerging technologies, Career pathways, Entrepreneurship, and Global Economy. The LinkedIn Economic Graph partners only with selected research teams and organisations. However, precise information on the requirements and selection criteria is not publicly available.

Source: Prognos AG/DevStat (2021).

After setting up the developer account it is possible to enable the required API products. With regards to the other official data access tools, precise information regarding the accessibility is not provided beforehand by LinkedIn.

3.2.4 Google Maps

The web mapping service Google Maps is a highly interesting platform that provides data for the construction of territorial indicators. In the following, the complex structure of the Google Maps data access possibilities is investigated.

³⁵ <https://docs.microsoft.com/en-us/linkedin/shared/authentication/getting-access?context=linkedin/context>

Table 3-5: Overview of Google Maps data access

Terms of use permit external data access	API available	Other official data access interfaces/tools available	No significant access restrictions
✓	✓	✗	✓

Source: Prognos AG/DevStat (2021).

With **three different APIs** (Maps API, Routes API, Places API), Google Maps provides a range of content that can be accessed. The Maps API provides access to Maps and Streetview, the Routes API can be used to receive directions and information about distances, and the Places API offers location data for over 100 million places. A highly complex structure is found in the regulations for the API rate limits and the pricing model which is highly intertwined.³⁶ The precise API rate limits depend on the type of API and the addressed endpoint. For some API calls, no rate limits are applied. Costs only occur when certain rate limits are exceeded. When having a credit of \$200 in the billing account most of the services are free of charge up to a certain amount of API requests.

Table 3-6: Factsheet Google Maps API

Google Maps API	
Available APIs	Maps API Routes API Places API
API-Content	<u>Maps API:</u> Maps, Street View <u>Routes API:</u> Directions, Distance Matrix, Roads <u>Places API:</u> Rich location data for over 100 million places. Enable to find places using phone numbers, addresses, and real-time signals
Regional granularity	NUTS 3
API rate limits	Different rate limits depending on web services (e.g. Daily quota starting at 100,000 requests per 24 hours, based on an annual contractual purchase. Maximum of 23 waypoints per request. Rate limit of 10 requests per second, etc.)
Access	User registration
Pricing	Some features are free. Prices depend on the type and number of calls. \$200 free monthly usage

Source: Prognos AG/DevStat (2021).

To **access the Google Maps APIs**, it is required to have and maintain a Google account in good standing. This rather vague access requirement is not specified by Google. Additionally, any information provided in

³⁶ <https://cloud.google.com/maps-platform/pricing/sheet>

the Google Account always must be accurate, correct, and up to date.³⁷ An application with a description of the use cases as required by other platforms is not necessary for the Google Maps API.

3.2.5 TripAdvisor

TripAdvisor is a tourism platform that provides both a booking & reservation system for hotels and transports as well as a recommendation platform. The options for accessing TripAdvisor data will be discussed in the following.

Table 3-7: Overview of TripAdvisor data access

Terms of use permit external data access	API available	Other official data access interfaces/tools available	No significant API access restrictions
✓	✓	✗	✗

Source: Prognos AG/DevStat (2021).

TripAdvisor offers access to its data through its Content API.³⁸ By using this method of data access it is possible to retrieve data such as locational data (e.g., on hotels, restaurants), reviews, ratings, and pricing of the hotels, restaurants, etc. listed on TripAdvisor. Regarding the API rate limits, TripAdvisor follows a two-step approach in which a limit of 50 calls/second and 1,000 calls per day is in place for developers using the API during the period of development and QA. Once the application is approved for launch, the daily limit increases to 10,000 calls.

To access the API, a request for API access must be filled out. It is highly important to mention that **TripAdvisor handles their API access highly restrictive**: *“TripAdvisor grants only a limited number of API keys and does not allow access to the Content API for purposes of Data analysis, Academic research”*.³⁹ This constitutes a significant restriction for the access to TripAdvisor data via API. Unfortunately, TripAdvisor does not provide more specific information regarding this issue.

Table 3-8: Factsheet Tripadvisor API

TripAdvisor API	
Available APIs	Content API
API-Content	Location data (name address, etc.) Reviews Ratings Prices
Regional granularity	NUTS 3 (Street level data)
API rate limits	Limit of 50 calls/second and 1,000 calls per day for developers using Content API during the period of development and QA. Once the application is approved for launch, the daily limit increases to 10,000 calls.

³⁷ <https://developers.google.com/maps/terms-20180207>

³⁸ <http://developer-tripadvisor.com/content-api/>

³⁹ <https://developer-tripadvisor.com/content-api/request-api-access/>

Access	Access needs to be requested. No access to Content API for purpose of data analysis and academic research.
Pricing	free





Source: Prognos AG/DevStat (2021).

Due to this significant access restriction to TripAdvisor data for data analysis and research an alternative data access needs to be considered. Alternatively, a web scraping approach can be followed.

3.2.6 ResearchGate

ResearchGate as a knowledge platform can provide (among others) information on scientific publications, citations, and authors. However, the platform does currently not provide an API. In the absence of other data access options **scraping ResearchGate is the second-best option for extracting data from the platform**. This access option will be explained in more detail in the following.

Table 3-9: Overview of ResearchGate data access

Terms of use permit external data access	API available	Other official data access interfaces/tools available	No significant API access restrictions
			

Source: Prognos AG/DevStat (2021).

As briefly mentioned previously a web crawler (or scraper) is a program that can be used to download web pages, extract links, and create lists of URLs. In other words, a web crawler is a tool that can be used to systematically extract web pages and their content. The most known example of a web crawler is the search engine developed by Google (Nemeslaki et al. 2011). Prognos has developed such a tool (*prognos web intelligence tool*) that is not available in the free market. This tool enables collecting and analysing information from the internet or other electronic sources. To determine which content of a website can be scraped it is crucial to examine the robot's exclusion protocol. As explained previously this protocol or robots.txt file states a standard in the communication between websites and web crawlers. It clarifies which pages or files a crawler can or cannot request from a website. In principle, the robot's exclusion protocol of ResearchGate⁴⁰ is open for web crawling.

3.2.7 Key takeaways for the framework of cooperation

The previous sections have shown that a platform's API is the method of choice for extracting big data on a regular basis. For all platforms except ResearchGate, an API is available and for most platforms on the short list, there are no significant access restrictions (except for TripAdvisor). Rate limits that curb the maximum amount of data that can be extracted in a given amount of time through an API are used by all platforms. A summary of the platform-specific findings is presented in the following table.

⁴⁰ <https://www.researchgate.net/robots.txt>

Table 3-10: Overview of API availability and restrictions

Platform	API available	API access restrictions	API rate limits
Twitter	Yes	No, if developer account available	Yes
LinkedIn	Yes	No, if developer account available	Yes
Google Maps	Yes	No, if developer account available	Yes
TripAdvisor	Yes	Yes	Yes
ResearchGate	No	n.a.	n.a.

Source: Prognos AG/DevStat (2021).

Core elements of the data access framework

Based on the previous findings the elements crucial to consider from a **technical perspective** are the following:

Table 3-11: Elements of the data access framework

Elements of the data access framework	Description
Terms of use permit data access	It needs to be checked whether a platform's terms of use do generally permit data access to external users
Platform offers data access via API	It needs to be checked whether a platform's grants access to its data via an API
No significant access restrictions of API / other official data access interface	It needs to be verified that there are no relevant restrictions that prohibit access to a platform's API / other official data access interface
No significant web scraping restrictions (if applicable)	It needs to be checked which data can be accessed through a web scraping approach
Platform-specific terms of use	It needs to be checked whether a platform prohibits the use of their data in a relevant way (i.e., in a way that is relevant for the construction of indicators)

Source: Prognos AG/DevStat (2021).

As it was outlined in section 3.1 next to the data access the processing of (personal) data constitutes a key area of tension. It is highly important that when processing such data an appropriate design is applied, for instance, through anonymisation or pseudonymization.

4 Territorial indicators based on big data

This chapter presents the approach for developing a list of territorial indicators. The proposed indicators are SMART indicators that are relevant for territorial policy-making and can be disaggregated in terms of well-defined small geographical areas (such as NUTS 3 areas, LAUs Level 1 or 2, or geographical grids) based on big data to be obtained from the private digital platforms selected in the Chapter 2 (Twitter, LinkedIn, TripAdvisor, and ResearchGate, as well as Google Maps). The rationale after the construction of the list of territorial indicators is led by diversity more than by exhaustivity, to explore the potentiality and limitation of the big data approach. In particular, the list includes both contextual indicators (measuring general trends and evolution of relevant topics) and policy indicators (aiming at measuring the level of achievement of selected policy goals). The list does not include impact indicators to be used in impact analysis of specific territorial policies, nor composite indicators.

Section 4.1 gives further insights on the georeferencing of territorial indicators, Section 4.2 outlines strategies to estimate territorial indicators and Section 4.3 illustrates the respective reference periods. Section 4.4 builds on the findings of Chapter 2 (selection and relevance of platforms) and shows a potential list of territorial indicators. The ones selected for the demonstrator study are described in more detail in Section 4.5.

4.1 Georeferencing of territorial indicators

As described previously data from the five selected platforms can be acquired using the corresponding APIs, web crawling procedures, or through legal agreements with the platforms. A critical point for the computation of territorial indicators, with any of the three strategies described below, is the geotagging of the information and its mapping into geographical areas to be relevant for territorial policy-making (such as NUTS 3 areas or LAU Level 1 and 2).

Geographical information from the platforms is geo-rereferred into two different ways:

- **Text geotag**, given by an address or, more commonly, just by the name of a location. For instance, Twitter's API provides the name of the geographical unit (municipality or country) of the user as a free text variable. After the corresponding editing procedure, the municipality or country can be mapped to NUTS areas using appropriate nomenclators from EUROSTAT⁴¹.
- **Geographical coordinates**, that can be mapped to NUTS/LAU or population grids using Eurostat's geographical information (GISCO⁴²). For instance, Twitter's API provides the coordinates of the location where each text has been tweeted.

4.2 Estimation strategies for territorial indicators based on data from private digital platforms

Proposed indicators can be computed using one of a combination of the following strategies:

- **Point-process estimation**: for numerical variables that are measured in a point with given geographic coordinates that to be directly downloaded from the platform, the value of the indicator for a geographical unit can be obtained just by aggregating the value of the variable in all the coordinates with each NUTS/LAU area. For instance, *the percentage of female recruitment processes in digitalisation related positions*. The aggregation can be done as the total number, proportion, mean. A more sophisticated approach could also define a territorial indicator from the spatial distribution of the observation within the NUTS/LAU.

⁴¹ <https://ec.europa.eu/eurostat/web/nuts/nuts-maps>

⁴² <https://ec.europa.eu/eurostat/web/gisco>

- **Distance-based estimation:** the value of the indicator is obtained as the aggregation of a pre-defined distance measure for all the cases within the unit. The distance can be defined through different approaches (geodesic distance: length of the shortest path connecting two points, shortest driving/walking/public transportation distance, shortest driving/walking/public transportation time required to travel to destination, etc). For instance, *the percentage of the population of a NUTS/LAU area who need more than 30 minutes to travel to a COVID19 vaccination centre*).

Box 4-1: Estimation with distances



Measuring different types of distances

While the geodesic distance between two points can be directly obtained from the coordinates of two points by GIS software, the proposed strategy is to use Google Maps (and its embedded travel optimisation systems) as the source for other relevant distance definitions. For trip duration estimations, that depend on the specific moment when they are computed, a reference time needs to be decided (for instance, a working day out of peak hours such as the first Tuesday of March at 10:30) or average values (average of the trips durations estimated each half an hour during a working day) should be specified.

Distance between a point (such as the location where a text has been twitted) and one-dimension (a highway) or two-dimension (a natural park) element can also be considered for the definition of territorial indicators.

Measuring population within a given distance

Relevant geographical units (such as a NUTS 2 area or the LAU corresponding to a large municipality) do not have a uniform distribution of population. Therefore, to determine how many people live within a 'ball' of a given 'radius' (for instance, population that needs to drive less than one hour to reach an airport), a higher granularity of population distribution is required. The project suggests using Eurostat's population grids⁴³, which provide the population in each cell of a square grid of 1 Km X 1 Km. For estimation purposes, the recommendation is considered all the population of each cell as living in its central point.

Source: Prognos AG/DevStat (2021).

- **Area estimation.** Supervised machine learning algorithms can be used to identify and measure the total size or percentage of the NUTS/LAU area with a given relevant feature in satellite images from Google Maps. For instance, this can allow for the estimation of *the percentage of land allocated to different uses* and its evolution over time, a critical point in the Territorial Agenda 2030.
- **Sentiment estimation.** Sentiment analysis refers to identifying as well as classifying the sentiments that are expressed in a text source. In a basic approach, sentiment can be classified as positive, neutral, or negative, although more sophisticated scales can also be considered, such as five levels Likert scales (very positive / positive / neutral / negative / very negative). Tweets are often useful in generating a vast amount of sentiment data upon analysis. After selecting a set of Tweets referred to a key concept or given hashtag, two main types of sentiment analysis solutions can be applied to classify each tweet as positive, neutral, or negative: (1) a rule system based in a

⁴³ <https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/geostat>

lexicon of words and rules and (2) the use of supervised machine learning predictive algorithms to be trained from a set of manually labelled tweets. Sentiment analysis allows for the definition of two types of indicators: although the total number or proportion of tweets on a topic with a positive sentiment provides information on the average attitude of the society on the NUTS/LAU towards the topics, the analysis of the distribution of the locations where positive, neutral, and negative sentiments are located provides information of the level of homogeneity of the society in each group. For instance, NUTS where around each half of the population exhibit positive and negative sentiments respectively (with no neutral comments), are highly polarised and may probably be prone to conflict.

- **Neighbourhood structure estimation:** Graph theory can be applied to the definition of connection indicators of geographical units. In this context, geographical units are referred to as vertices, or nodes, have varying connections between themselves referred to as edges or links determined by the corresponding flow territorial indicators. The **centrality degree** of a region can be defined as the number of direct connections a region has with others. Degree centrality indicators assign a score based simply on the number of links held by each region: the higher the degree of a node, the more important it is. The degree can be extended to the sum of weights when analyzing weighted networks.⁴⁴

4.3 Reference period and breakdown of territorial indicators

The estimation of a territorial indicator will be always referred to a specific time and population group.

- **Reference period** is the time period for which the estimator is estimated. It can be defined as a period of time (percentage of recruitment processes in a given specific field during the last month) or a date (percentage of workers in a specific field on December 31st, 2020). It needs to be specified in the definition of the indicator.
- **Profiling of indicators.** A critical strength of big data indicators is the detailed granularity of the source, which allows for breakdowns of estimators of specific groups of the population. However, the information on the users that can be acquired from APIs is quite limited, as shown in the case of Twitter in the table below. In this case, the information even for basic classification variables (such as gender, age, ethnicity, etc.) is not directly provided by the platform. Therefore, the classification variables to be used for the breakdowns should also be estimated, using supervised and unsupervised machine learning classification methods. For instance, these methods could help to classify the gender of a user not only from the profile of the user but for all the information in the platform, such as the complete texts of her/his tweets. This approach opens interesting possibilities for the establishment of sophisticated breakdowns, for instance, non-binary gender classifications, that cannot be achieved using classical sources.

⁴⁴ Opsahl, T., Agneessens, F., Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks* 32, 245-251

Table 4-1: Information on Twitter's user provided by the API

user_id	User identification
screen_name	User screen name
Name	Username
account_lang	Language of user account
account_created_at	Date of account creation
location	The user-defined location for this account's profile. Not necessarily a location.
profile_url	Profile URL
url	A URL provided by the user in association with their profile.
profile_image_url	Image uploaded by the user
followers_count	The number of followers this account currently has
friends_count	The number of users this account is following
listed_count	The number of public lists that this user is a member of
favourites_count	The number of Tweets this user has liked in the account's lifetime
protected	When true, indicates that this user has chosen to protect their Tweets
verified	When true, indicates that the user has a verified account

Source: Prognos AG/DevStat (2021).

4.4 Overview of potential territorial indicators

A list of potential indicators is shown in the table below. Without any intent of exhaustivity, the list presents a series of heterogeneous indicators that leverages the main strengths of big data and explores the different methods to extract information from big data (analysis of structured numeric variables, textual analysis, and image analysis).

Table 4-2: Overview of potential territorial indicators (selected examples, non-exhaustive)

Topics of the Territorial Agenda	Twitter	LinkedIn	TripAdvisor	ResearchGate	Google Maps
Quality of life	Sentiment indicator on happiness Sentiment indicator on EC policy Sentiment indicator on security Sentiment indicators urban environment (mobility, noise, illumination)) Sentiment indicator on LGBTBI+ issues	/	Average rating of leisure services (restaurants, attractions, etc)	/	/
Services of general interest	Sentiment indicator on housing (rent/purchase)	/	/	/	% of population within a given distance to (basic service) Average distance to (basic service) % of population within a given travel duration (foot/car/public) to (basic service) Average travel duration (foot/car/public) to (basic service)
Demographic and social imbalances	% of active female twitter users % of active twitter users from other relevant population groups % of texts twitted by women % of texts twitted by other relevant population groups (EU migration)	% of active female LinkedIn users % of active LinkedIn users in other relevant population groups % of workers by gender/specific group per field	Average rating of leisure and culture services by gender / groups of population	% of papers published by gender in relevant fields % of papers published with a woman as principal author in relevant fields Number of published papers on gender Number of published papers on intra/extra EU migration	% of population of a specific group within a given distance to (basic service) % of population of a specific group within a given travel duration (foot/car/public) to (basic service)
Digitalisation and 4th industrial revolution	% of population using Twitter Number of texts twitted/retweeted per 100,000 inhabitants	% of population using a LinkedIn Number of job offers requiring technical skills for digital innovation Number of work positions related to digital innovation Number of posts on digitalisation-related issues	% of tourist businesses (restaurants/hotels) with presence in TripAdvisor	Number of published papers on digitalisation related topics	Number of businesses from sector in danger from large digital companies (such as bookshops)
Employment and economic development	Sentiment indicator of confidence in the economy	Job offers (total / per 100,000) inhabitants by type of job offer and sector Workers (total / per 100,000) inhabitants by specific sectors	/	Number of published papers on I+D related issues	/

Interdependence between places	/	/	% of population traveling from/ to the area Travelling centrality level	/	/
Global embeddedness	Sentiment indicator on intra/extra EU migration	/	/	% of papers with EU and non-EU co-authors	% of population within distance to an airport % of population within a given travel duration (car/public) to an airport
Climate change	Sentiment indicator on urgency of climate change Number of tweets related to heat, floods etc.	/	/	Number of published papers on climate change related issues	
Loss of biodiversity, land consumption	Sentiment indicator on biodiversity loss (general and specific species)	/	/	Number of published papers on loss of diversity related issues	% of land covered by use (edification, green areas, freshwater surfaces, etc.)
Air, soil, water	/	/	/	/	/
Secure, affordable, and sustainable energy	Sentiment indicator on energy affordability	Job offers on renewable energies related activities Works on renewable energies related activities		Number of published papers on energy market and renewable energies	Number of renewable energy plants
Just transition	Sentiment analysis on the economic/well-being, impact of green transition	/	/	/	/
Circular value chains	/	/	/	/	/
Nature, landscape and cultural heritage	Sentiment indicator on visits to nature and cultural landmarks	/	Number of nature, landscape and culture landmarks Average rating on nature, and culture landmarks	Number of publications on nature, landscape and culture	/
COVID 19	Number of users/tweets denying pandemic Number of users/tweets denying COVID19 vaccination effectivity Sentiment indicators on the EC management of COVID19 crisis Sentiment indicator on pandemic manag. (EC) Sentiment indicator on pandemic management by national/local governments Sentiment indicator on teleworking	Number of recruiting processes in COVID-related jobs	/	Number of publications on COVID 19 related topics Centrality level in the co-authors network on COVID 19 related papers	Average distance to COVID vaccination centre % of population within a given distance to a hospital / ICU % of population within a given travel duration (foot/car/public) to a hospital / ICU

Source: Prognos AG/DevStat (2021).

Box 4-2: Generic indicators

i

Defining indicators to be customized to all the topics of the Territorial Agenda

Some of the indicators defined in this section can be easily extended to any other point of the territorial agenda, just by adapting the specific field for which the indicator has been estimated. This can be done for:

- *Saliency, sentiment, and polarity indicators* from Twitter. This family of indicators can measure the awareness and attitude towards any specific topics related to the territorial agenda among a subgroup of the population living in a NUTS/LAU area
 - Saliency indicators can be obtained as the total or relative (for instance, per 100,000 inhabitants) number of tweets mentioning the topic or using a related hashtag.
 - Sentiment indicators can be built from classifying tweets referring to any topic in the territorial agenda using rule systems or supervised machine learning algorithms.
 - Polarity indicators can be built from the distribution of positive, neutral, and negative tweets in a NUTS/LAU. High polarization can be related to those situations with a large percentage of tweets with positive/negative sentiment and a low percentage of neutral tweets.
- *Job market indicators* from LinkedIn. Using LinkedIn as a source, the total and relative (for instance, per 100,000 inhabitants) number of job offers and workers in specific fields related to the points in the territorial agenda can be used as a measure of the importance of any action points of the territorial agenda in an area.
- *Publication indicators* from ResearchGate. Using ResearchGate as a source, the number of publications in specific fields related to the points in the territorial agenda can be used as a measure of the importance of the different action points of the territorial agenda in an area.

Source: Prognos AG/DevStat (2021).

4.5 Selected indicators for the demonstrator study

A selection of indicators to be addressed in the demonstrator study has been made from the above list (Table 4-2). The demonstrator study follows two objectives. First, to assess the feasibility of big data acquisition from the private platforms and computation of big data territorial indicators. Second, to provide ESPON with validated information on the requirements to implement a massive estimation of big data territorial indicators and guidelines on good practices for this estimation. Against this background the selection criteria for the indicators in the demonstrator study will be described below:

- **Criterion 1: Heterogeneity.** To optimise the resources of the project and assess the feasibility and potentiality of the different types of possible indicators, the proposed indicators have been selected to cover as many sources and estimation approaches as possible.
- **Criterion 2: Data availability.** At this point of the project, where no commercial agreement has been already subscribed with the private platforms, the proposed indicators have been selected to be computed using public APIs provided with the platforms, web scraping, or manual downloading of specific examples of datasets.

- **Criterion 3: Relevance.** The short list indicators have been selected to cover relevant policy-making topics such as the COVID crisis, access to services, land use, and R&D.

The following tables give an **overview of the relevant indicator groups** that were chosen to be relevant for the demonstrator study. The respective indicators for each group that have been looked at in the demonstrator study are marked in red.

Table 4-3: Indicators relevant to the perception of COVID-19 crisis

TI1	Perception of the political management of Covid-19 crisis ⁴⁵
Definition	Citizens' perception of the way in which the EC, national and local governments have managed COVID19 pandemic in general and selected specific issues (border closing, centralized vaccine purchasing, vaccination effectivity denial, mobility policies such as temporary cycling infrastructures, etc.)
Set of possible indicators (selected)	<ul style="list-style-type: none"> • TI1_1: Number of tweets referring to the global management / specific issue by the EC, National, and Local governments (salience). • TI1_2: Number of retweets referring to the global management / specific issue by the EC, National, and Local governments (salience). • TI1_3: Number of retweets referring to the global management / specific issue by the EC, National and Local governments by area from where they were re-tweeted. • TI1_4: Number of users twitting texts referring to the global management / specific issue by the EC, National, and Local governments (salience). • TI1_5: % of population twitting texts referring to the global management / specific issue by the EC, National and Local governments from each other area. • TI1_6: % of positive tweets referring to the global management / specific issue by the EC, National, and Local governments (sentiment). • TI1_7: % of negative tweets referring to the global management / specific issue by the EC, National, and Local governments (sentiment). • TI1_8: % of neutral tweets referring to the global management / specific issue by the EC, National, and Local governments (sentiment). • TI1_9: Level of polarity of tweets referring to the global management / specific issue by the EC, National, and Local governments. • TI1_10: % of negative tweets referring to the global management / specific issue and living at less than 30 km to a hospital with ICU by the EC, National, and Local governments. • TI1_11: % of negative tweets referring to the global management / specific issue and living at more than 30 minutes driving from a hospital with ICU by the EC, National and Local governments. • TI1_12: Number of tweets on the topic covid [see demonstrator study in Chapter 5] • TI1_13: Number of tweets on the topic covid per inhabitants [see demonstrator study in Chapter 5] • TI1_14: Number of tweets on vaccines from capital cities per 10,000 inhabitants [see demonstrator study in Chapter 5]
Sources	<ul style="list-style-type: none"> • Twitter: tweets for the demonstrator study acquired from twitter's public API. • Google Maps: location of hospitals with ICU to be manually acquired

⁴⁵ The results of the ESPON research project *Territorial impacts of Covid-19 and policy answers in European regions and cities* (<https://www.espon.eu/covid-19>) could provide guidance to define the specific issues of policy management of Covid-19 crisis to be used in the definition of potential subindicators under IT1.

	<ul style="list-style-type: none"> • Eurostat population grid, to be downloaded from GISCO website⁴⁶
Computation	<p>The tweets referring to the global management or to specific issues to be evaluated can be selected and classified as positive, neutral, or negative by using deterministic rules from a predefined lexicon of words. The demonstrator study considers only tweets in English for an area to be defined in Ireland. The location of each tweet is provided by its geographical coordinates, available through the API. The location of the user can be obtained after mapping the name of the location provided by the API in the corresponding NUTS/LAU area using appropriate nomenclators from Eurostat.⁴⁷</p> <ul style="list-style-type: none"> • TI1_1 to TI_4 and TI1_6 to TI_8 are obtained by aggregation of the downloaded tweets, once classified as positive, neutral, or negative. • TI1_5 is obtained as the ratio of TI1_4 and the total population of the area to which the indicators refer, obtained by aggregating the population data in Eurostat's population grid. • TI1_9 is obtained as $(TI1_6) \times (TI1_7) / 2500$. The indicator takes values from 0 (minimum polarity) and 1 (maximum polarity when $TI1_6 = TI1_8 = 50\%$). • TI1_10 and TI1_11 are estimated by determining the distance from the coordinates of the twitting location and the closest hospital with ICU and computing the % of those tweets classified as negative <p>The reference period is the week before the date when the tweets have been acquired.</p>
Breakdowns	<ul style="list-style-type: none"> • NUTS3 and LAU level 2 • Gender, to be determined using a supervised ML algorithm using the available information of the user
Associated second-tier indicators	<ul style="list-style-type: none"> • Daily time series of the evolution of the basic indicators • Ranking of areas in terms of salience, sentiment, and polarity of the basic indicators • Degree of retweeting centrality of each NUTS3 / LAU level 2 area in terms of TI1_5 • Comparison of the perception of the management of the COVID19 crisis by different levels of government.

Source: Prognos AG/DevStat (2021).

Table 4-4: Indicators relevant for technical skills for digital innovation

TI5	Supply and demand of technical skills for digital innovation
Definition	Needs and availability of human resources with the skills required to implement the digital transformation in the public and private sectors.
Sub-indicators	<ul style="list-style-type: none"> • TI5_1: Number of job offers requiring technical skills for digital innovation • TI5_2: % of job offers requiring technical skills for digital innovation • TI5_3: Number of workers in positions related to digital innovation • TI5_4: % of workers in positions related to digital innovation • TI5_5: Number of students of technical education programs related to digital innovation

⁴⁶ <https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/geostat>

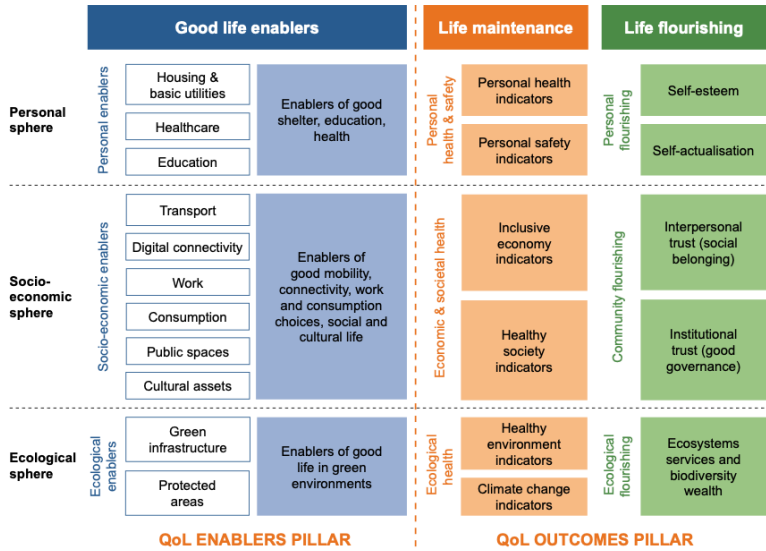
⁴⁷ <https://ec.europa.eu/eurostat/web/nuts/nuts-maps>

	<ul style="list-style-type: none"> • TI5_6: % of students of technical education programs related to digital innovation • TI5_7: Total workers with programming skills [see demonstrator study in Chapter 5] • TI5_8: % of workers with programming skills in the total LinkedIn Members [see demonstrator study in Chapter 5] • TI5_9: Total workers with skills in the field of AI [see demonstrator study in Chapter 5] • TI5_10: % of workers with skills in the field of AI in the total LinkedIn [see demonstrator study in Chapter 5] • TI5_11: Total workers with skills in the field of Robotics [see demonstrator study in Chapter 5] • TI5_12: % of workers with skills in the field of Robotics in the total LinkedIn Members [see demonstrator study in Chapter 5] • TI5_13: Total number of LinkedIn Members [see demonstrator study in Chapter 5]
Sources	<ul style="list-style-type: none"> • LinkedIn, data to be acquired using web scrapping or small-scale manual downloading.
Computation	Lexicons of words referring to technical skills for digital innovation are defined in different languages. The job offers for TI5_1 and TI5_2 can be identified as those offers including the word in these lexicons in the job description. In a similar way, lexicons of words associated with positions (education programs) related to digital transformation are defined in the different languages. These lexicons are used to identify the workers and students in these areas through the job description in the user profile. After the assignation of each researcher, offer, and work to its NUTS/LAU area, the total number and % in TI5_1 to TI5_6 can be computed by aggregating the individual cases in each area.
Break-downs	<ul style="list-style-type: none"> • NUTS3 and LAU level 2 • Public⁴⁸ and private entities publishing the job offer or where the user is working.
Associated second-tier indicators	<ul style="list-style-type: none"> • Annual time series of the evolution of the basic indicators • Correlation between supply and demand of skills for digitalisation • Degree of research centrality of each NUTS3 / LAU level 2 area in terms of TI4_3

Source: Prognos AG/DevStat (2021).

⁴⁸ Using this breakdown, the indicators could complete the information provided by the massive survey in the DIGISER research project. Specifically, IT5 will allow for the analysis of the supply and demand of technical skills to deploy digital transformation and how they are allocated to the public and private sector.

Table 4-5: Indicators relevant for quality of life

Tl6	Citizen perception of the enablers and outputs of territorial quality of life
Definition	<p>Citizens' perception on the 11 good life enablers and 11 good life outputs proposed in ESPON (2021) working paper on quality-of-life measurement, as presented in the following figure:</p>  <p style="text-align: center;"><i>Source: ESPON (2021) Is our life good enough?⁴⁹</i></p>
Sub-indicators	<p>For each of the 11 enablers and 11 outputs of good life in ESPON (2021):</p> <ul style="list-style-type: none"> • Tl6_1: Number of tweets referring to each enabler / output (salience). • Tl6_2: Number of retweets referring to each enabler / output (salience). • Tl6_3: Number of retweets referring to each enabler / output by area from where they were re-tweeted. • Tl6_4: Number of users twitting texts referring to each enabler / output (salience). • Tl6_5: % of population twitting texts referring to each enabler / output. • Tl6_7: % of positive tweets referring to each enabler / output (sentiment). • Tl6_8: % of negative tweets referring to each enabler / output (sentiment). • Tl6_9: % of neutral tweets referring to each enabler / output (sentiment). • Tl6_10: Level of polarity of tweets referring to each enabler / output.
Sources	<ul style="list-style-type: none"> • Twitter: tweets for the demonstrator study are acquired from twitter's public API. • Eurostat population grid, to be downloaded from GISCO website⁵⁰
Computation	<p>The tweets referring to each specific good life enabler / output to be evaluated can be selected and classified as positive, neutral, or negative by using deterministic rules from a predefine lexicon of words. The demonstrator study considers only tweets in English for an area to be defined in Ireland.</p>

⁴⁹ <https://www.espon.eu/is-our-life-good-enough>

⁵⁰ <https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/geostat>

	<p>The location of each tweet is provided by its geographical coordinates, available through the API. The location of the user can be obtained after mapping the name of the location provided by the API in the corresponding NUTS/LAU area using appropriate nomenclators from Eurostat.⁵¹</p> <p>TI6_1 to TI6_4 and TI6_6 to TI_8 are obtained by aggregation of the downloaded tweets, once classified as positive, neutral or negative.</p> <ul style="list-style-type: none"> • TI6_5 is obtained as the ratio of TI6_4 and the total population of the area to which the indicator refers, obtained by aggregating the population data in Eurostat's population grid. • TI6_9 is obtained as $(TI6_6) \times (TI6_7) / 2500$. The indicator takes values from 0 (minimum polarity) and 1 (maximum polarity when $TI6_6 = TI6_8 = 50\%$). <p>The reference period is the week before to the date when the tweets have been acquired.</p>
Break-downs	<ul style="list-style-type: none"> • NUTS3 and LAU level 2 • Gender, to be determined using a supervised ML algorithm using the available information of the user
Associated second-tier indicators	<ul style="list-style-type: none"> • Daily time series of the evolution of the basic indicators • Ranking of areas in terms of salience, sentiment, and polarity of the basic indicators • Degree of retweeting centrality of each NUTS3 / LAU level 2 area in terms of TI6_3

Source: Prognos AG/DevStat (2021).

Other groups of indicators that were further looked into but not used for the demonstrator study are illustrated in the following tables.

Table 4-6: Indicators relevant to access to bank services

TI2	Access to offline bank services
Definition	Population that has access to financial services in bank offices or ATMs
Sub-indicators	<ul style="list-style-type: none"> • TI2_1: % of population living less than 1 km from a bank office • TI2_2: % of population living less than 15 minutes walking from a bank office • TI2_3: % of population living less than 30 minutes driving from a bank office • TI2_4: % of population living less than 1 km from an ATM • TI2_5: % of population living less than 15 minutes from an ATM • TI2_6: % of population living less than 30 minutes driving from an ATM
Sources	<ul style="list-style-type: none"> • Google Maps: location of bank offices and ATMs, to be manually acquired • Eurostat population grid, to be downloaded from GISCO website⁵²
Computation	<p>For the computation of these indicators, one can assume that the population living in each 1km x 1km cell of Eurostat's population grid is living in a point located at the centre of the cell. Walking and driving travel duration can be estimated using Google Maps at a fixed time (for instance, at 12:00 on the first Wednesday of the month).</p> <p>TI2_1 to TI2_6 can be computed as the % of the aggregated population from Eurostat's grid in the area that lies in a ball centred at the closest bank office / ATM and a radius given by the distance / travel time specified in the definition of the corresponding indicator.</p>

⁵¹ <https://ec.europa.eu/eurostat/web/nuts/nuts-maps>

⁵² <https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/geostat>

Break-downs	<ul style="list-style-type: none"> • NUTS3 and LAU level 2 • Gender, to be determined using a supervised ML algorithm using the available information of the user
Associated second-tier indicators	<ul style="list-style-type: none"> • Annual time series of the evolution of the basic indicators • Ranking of areas in terms of offline access to financial services

Source: Prognos AG/DevStat (2021).

Table 4-7: Indicators relevant to land use

TI3	Land use
Definition	Land allocated to each potential uses (edifications, green areas, crops, etc.)
Sub-indicators	<ul style="list-style-type: none"> • TI3_1: Number of Km2 allocated to each land use • TI3_2: % of surface allocated to each land use • TI3_3: % of population living less than 1 km from a green area • TI3_4: % of population living less than 15 minutes from a green area • TI3_5: % of population living less than 30 minutes driving from a green area
Sources	<ul style="list-style-type: none"> • Google Maps: satellite images of the area • Eurostat population grid, to be downloaded from GISCO website⁵³
Computation	<p>TI3_1 and TI3_2 can be computed by identifying and classifying areas of different uses using a supervised machine learning algorithm of image recognition. The algorithm can be trained after manually contouring and labelling areas with different uses in a training imagen dataset obtained from Google Maps.</p> <p>For the computation of TI3_3 to TI3_5, one can assume that the population living in each 1km x 1km cell of Eurostat's population grid is living in a point located at the centre of the cell. Walking and driving travel duration can be estimated using Google Maps at a fixed time (for instance, at 12:00 of the first Wednesday of the month). The indicator can then be computed as the % of the aggregated population from Eurostat's grid in the area within distance / travel time to the closest green are specified in the definition of the corresponding indicator.</p>
Break-downs	<ul style="list-style-type: none"> • NUTS3 and LAU level 2 • Type of crop, to be determined using a supervised ML algorithm.
Associated second-tier indicators	<ul style="list-style-type: none"> • Annual time series of the evolution of the basic indicators • Ranking of areas in terms of land use

Source: Prognos AG/DevStat (2021).

⁵³ <https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/geostat>

Table 4-8: Indicators relevant to R&D employment territorial indicators

TI4	R&D – employment territorial indicators
Definition	Level and specialization of R&D activities in the area
Sub-indicators	<ul style="list-style-type: none"> • TI4_1: number of publications in a specific research field • TI4_2: % of publications in a specific research field • TI4_3: number of publications in a specific research field co-authored with researchers of other areas • TI4_4: number of job offers in a specific activity field • TI4_5: % of job offers in a specific activity field • TI4_6: number works in a specific activity field • TI4_7: % of works in a specific activity field
Sources	<ul style="list-style-type: none"> • LinkedIn, data to be acquired using web scrapping or small-scale manual downloading. • ResearchGate, data to be acquired using web scrapping or small-scale manual downloading
Computation	<p>The location of the affiliation organization (university) of the authors of a publication, the location of the company offering a job, and the location of each worker can be obtained after mapping the name of the location to be obtained from the platforms in the corresponding NUTS/LAU area using appropriate nomenclators from Eurostat.⁵⁴</p> <p>After the assignation of each researcher, offer, and work to its NUTS/LAU area, the total number and % in TI4_1 to TI4_6 can be computed by aggregating the individual cases in each area.</p>
Break-downs	<ul style="list-style-type: none"> • NUTS3 and LAU level 2 • Research and activity fields referred to as ISIC rev4 sectors with the maximum granularity
Associated second-tier indicators	<ul style="list-style-type: none"> • Annual time series of the evolution of the basic indicators • Spatial correlation of research activities and economy activities • Correlation between research intensity and economic activity • Degree of research centrality of each NUTS3 / LAU level 2 area in terms of TI4_3

Source: Prognos AG/DevStat (2021).

4.6 Going beyond the basic territorial indicators: Second-tier indicators

The indicators defined in the previous subsections can be manipulated to create in the future second-tier indicators measuring the evolution, heterogeneity, and convergence of the underlying topic in the different NUTS/LAU areas:

- **Trend indicators.** Once an indicator has been estimated at two reference dates, the percentage of variation from the former to the latter date can be used to measure its evolution over time.
- **Comparison indicators.** A comparison of the estimation of the indicators among the different regions allows for building (1) ranking of regions, (2) quantification of the difference in the concept

⁵⁴ <https://ec.europa.eu/eurostat/web/nuts/nuts-maps>

measured by the indicator among regions and with respect to the average value in the EU or any higher-level area (NUTS0) and (3) measures of inequality among regions.

- **Convergence indicators.** Using a combination of trend/comparison approaches, it is possible to evaluate if different regions converge/diverge according to the key topics of the Territorial Agenda 2030.
- **Nowcasting indicators.** Statistical models can be developed to relate high-frequency (even real-time) basic territorial indicators based on big data with classical low-frequency indicators to provide nowcasting.
- **Composite indicators.** Both big data and conventional territorial indicators can be used as building blocks to define composite indicators capable to measure general topics in the territorial agenda or the global territorial trends and impacts of territorial policies.

Box 4-3: Spatial data analysis

i

The Nobel Prize winner Paul Krugman, father of economic geography, describes economic geography as⁵⁵ "that branch of economics that worries about where things happen in relation to one another". Translating this concept to more general territorial policymaking, the computation of the territorial indicators with a breakdown at NUTS/LAU area level, as defined in previously, is just the first step to provide information to define and evaluate policies for the action point of the Territorial Agenda 2030. To this end, spatial analysis starts by **describing the location of the territorial indicators**, then **measuring the importance of spatial interactions** in order to be able to take these interactions among territorial indicators into account using an appropriate model.

One of the characteristics of spatial analysis is that the spatial coordinates contain potentially meaningful information. To make use of such information, the first step is to group the data according to their geographical proximity. The second stage of spatial analysis consists in defining a NUTS/LAU neighbourhood. Defining the neighbourhood is an essential step toward measuring the strength of spatial relationships between NUTS/LAU, in other words, the way in which neighbours influence each other. The more the observation values are influenced by values of observations that are geographically close to them, the greater the spatial autocorrelation. Spatial autocorrelation indices measure the strength of spatial interactions between observations. Spatial econometrics models this spatial dependency. There are multiple forms of interactions related to the variable to be explained, the explanatory variables or the unobserved variables. As a result, these many models end up in competition, all building from the same prior definition of neighbourhood relations. Spatial heterogeneity refers to the fact that the influence of explanatory variables on the dependent variable varies with the location of the observations. Geographically weighted regression or spatial smoothing are used to take this phenomenon into account⁵⁶.

Source: Prognos AG/DevStat (2021).

⁵⁵ Krugman, Paul R (1991). Geography and trade. MIT press.

⁵⁶ Loonis, V., & de Bellefon, M. P. (2018). Handbook of Spatial Analysis: Theory and Application with R. Eurostat, INSEE, no. October, 394. Available at <https://www.insee.fr/en/information/3635545>

5 Demonstrator Study: Territorial big data

Based on the defined set of territorial indicators, selected in 4.5, this chapter aims to describe the **methodology and approach of how big data can be utilized for territorial analysis** and point out the main added value such analysis can provide to policymakers.

For this purpose, the following **three topics** with high relevance in the Territorial Agenda 2030 and current policy making were selected in coordination with ESPON:

1. Technical Skills for Digital Innovation
2. COVID-19

Moreover, to explore these topics with data from private digital platforms, **two relevant platforms** were selected, namely LinkedIn and Twitter. LinkedIn is a platform for professionals and contains data on job profiles, CVs and skills of its members. It is therefore well suited for modelling indicators on the topic of Technical Skills for Digital Innovation. To demonstrate new possibilities of measurement data from Twitter can be used. Twitter makes it possible to show in which regions and to what extent topics occur in the public debate. Due to the huge amount of data, a few regions were selected for the study to demonstrate the approach. For the selected regions, data was collected and analysed according to the selected indicators. The procedure is described in the chapters on methodology. The results of the analysis are then presented in the chapters Illustrations of findings.

5.1 Selection of pilot regions

To arrive at a 'proof of concept' of our approach on the NUTS 3 level, different types of regions, i.e., both rural and urban areas as well as the European linguistic diversity, must be taken into account. Therefore, three different region types at NUTS 3 level in each of four different countries have been selected. Countries with different languages and different population sizes have been selected. Table 5-1 shows the selected regions.

Table 5-1: Selected Regions for pilot study

Country	NUTS Code	Regions	Main language	Population ⁵⁷
Germany	DE	Germany	German	83.166.711
Germany	DE300	Berlin	German	3.669.491
Germany	DE212	Munich	German	1.484.226
Germany	DE80L	Vorpommern-Rügen	German	224.702
Ireland	IE	Ireland	English	4.964.440
Ireland	IE061	Dublin	English	1.408.815
Ireland	IE053	South-West Ireland	English	712.968
Ireland	IE042	Western Ireland	English	468.945
Poland	PL	Poland	Polish	37.958.138
Poland	PL911	Miasto Warszawa	Polish	1.789.771
Poland	PL213	Kraków	Polish	775.654
Poland	PL721	Kielce County (Kielecki)	Polish	749.532
Spain	ES	Spain	Spanish	47.332.614
Spain	ES30	Madrid	Spanish	6.747.068
Spain	ES618	Seville	Spanish	1.957.520

⁵⁷ <https://ec.europa.eu/eurostat/web/main/data/database> (last update: 11/03/2021)

Country	NUTS Code	Regions	Main language	Population ⁵⁷
Spain	ES212	Guipúzcoa	Spanish	716.552

Source: Prognos AG/DevStat (2021).

In the next sections, these selected regions will be used to explore the usage of territorial big data from private platforms.

5.2 LinkedIn: Technical skills for digital innovation

To assess 'technical skills for digital innovation' in the selected pilot regions based on LinkedIn data, two different indicators have been chosen:

- TI5_7: Total workers with programming skills
- TI5_8: % of workers with programming skills in the total LinkedIn Members
- TI5_9: Total workers with skills in the field of AI
- TI5_10: % of workers with skills in the field of AI in the total LinkedIn
- TI5_11: Total workers with skills in the field of Robotics
- TI5_12: % of workers with skills in the field of Robotics in the total LinkedIn Members
- TI5_13: Total number of LinkedIn Members

Below we describe the methodology and illustrative findings from this analysis, including a comparison with public statistics (where possible).

5.2.1 Methodology

LinkedIn provides different options to access information on its platform. The selected indicators *TI5_4: % of workers in positions related to digital innovation (defined and specified as workers with programming skills)* and *TI5_7: % of LinkedIn Members with skills in the field of AI and Robotics in the total LinkedIn Members*, require access to information about people's skills in different regions. For this purpose, the Tool 'Recruiter Lite' from the LinkedIn Talent Solutions Tool family can be used. The data has been collected on 30 August 2021 and was used as point-in-time data.

The Recruiter Lite Tool offers a search mask with different filter functions e.g., position, location, skills, company, final year, training, or industry. To access relevant data, the location and skills filters have been used. Although only 1st-degree, 2nd-degree, and 3rd-degree connections can be viewed and contacted, higher-degree connections appear in the Recruiter Lite tool search results with the member's headline, location, and years of experience.⁵⁸ To verify that the extracted numbers approximate LinkedIn's user base, user statistics published by LinkedIn can be referenced.⁵⁹ A comparison of the extracted figures for Poland (3.75 million) and Spain (14.09 million) with the rounded user base figures published by LinkedIn for Poland (4 million) and Spain (14 million) confirms the convergence of the extracted figures with the actual user base. No user figures were published for the countries Germany and Ireland and can therefore not be compared. Although the figures converge, it must be emphasized at this point that a bias may exist due to the use of a specific network.

For each region, the ontology from Table 5-2 (see below) was entered into the skills filter and the corresponding number of candidates was extracted. The ontology on AI and Robotics is based on *Keywords used*

⁵⁸ Recruiter Lite Profile Visibility and Messaging Capabilities: <https://www.linkedin.com/help/recruiter/answer/a414468>

⁵⁹ <https://news.linkedin.com/about-us#>

in LinkedIn queries from the *Advanced Technologies for Industry – Methodological report*⁶⁰ and was enriched with the different languages where relevant. Additionally, the field 'programming' was added. It can be used to approximate the order of magnitude of the proportion of LinkedIn members with in-depth digital skills as there is a link between programming skills and in-depth digital skills. Furthermore, skills in the field of AI are much more specific than skills in the field of programming and therefore it is expected that the share of members with skills in the field of AI is smaller.

Table 5-2: Ontology for selected technology areas

Technology field	Keywords
AI	artificial intelligence; biometrics; cognitive computing; computer vision; deep learning; machine learning; natural language processing; Natural language understanding; natural language generation; reinforcement learning; speech recognition; supervised learning; unsupervised learning; neural network; Künstliche Intelligenz; Maschinelles Lernen; Neuronale Netzwerke; sztuczna inteligencja; Uczenie maszynowe; Sieci neuronowe; inteligencia artificial; Aprendizaje automático; Redes neuronales;
Robotics	robotics; robot; human robot interfaces; Robotic; drones; robotic surgery; robotic human interaction; Robotik; Roboter Chirurgie
Programming	JavaScript; Java; PHP; C++; Python; SQL; HTML; CSS; MATLAB; R ; Programming; Web Development ; App Development; Digital Design; Data Visualization; user interface; Power BI; User Experience Design (UX); Data Analytics; Data science

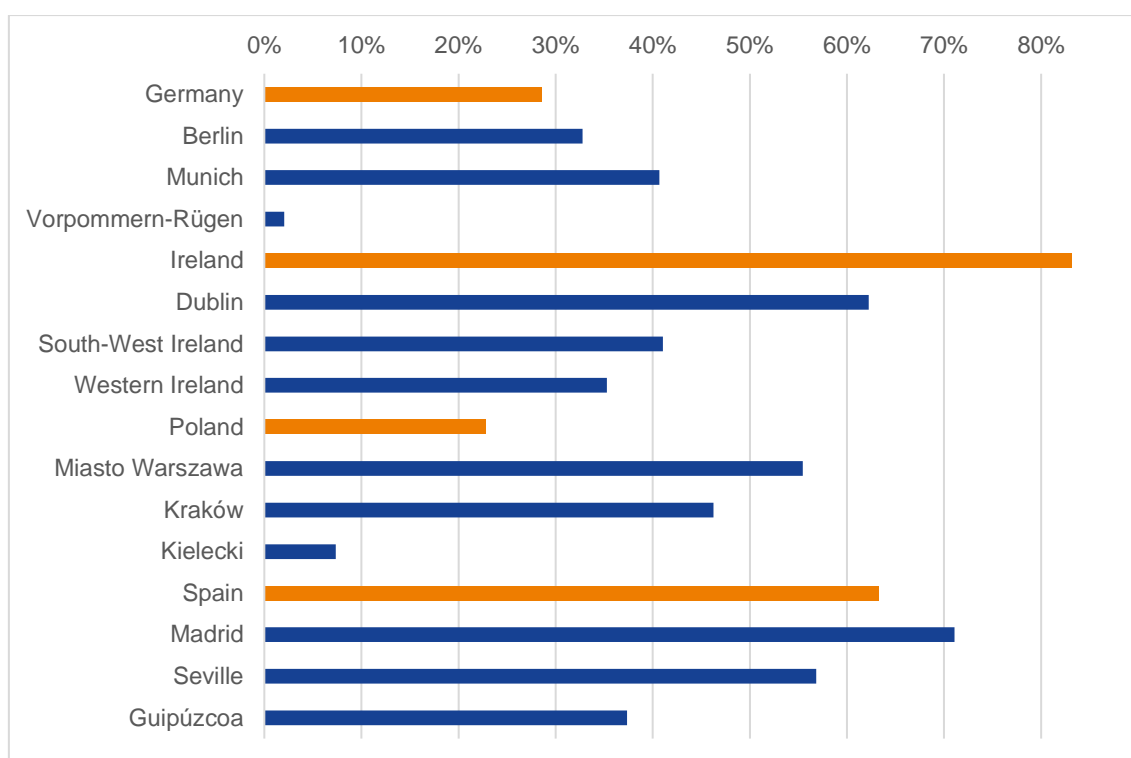
Source: Prognos AG/DevStat (2021).

Since the query of the different regions on LinkedIn is based on the same ontology, i.e., the same keywords, the regions are comparable with each other to the extent that the same skills and qualifications were queried. One limitation is the different share of LinkedIn Members in the employed population in the regions as shown in Figure 5-1. The share of LinkedIn members in the employed population in Ireland (~83%) and Spain (~63%) is higher than the share in Germany (~28%) and Poland (~23%). Moreover, the share of LinkedIn members in the total population is higher in urban areas than in rural regions. Employment data were extracted from the Eurostat database (Employment by NUTS 3 regions) and are from 2018 (most recent).

In addition, the average share of LinkedIn members in the employed population at a country level is in most cases lower than in urban regions. Ireland is an exception. This gap between Ireland as a whole country and the different Irish regions might be explained by Irish LinkedIn Members that provide Ireland as their home region without specifying a region in LinkedIn.

⁶⁰ European Commission (2020): Advanced Technologies for Industry - Methodological report. <https://ati.ec.europa.eu/reports/eu-reports/advanced-technologies-industry-methodological-report>

Figure 5-1: Share of LinkedIn Members in the employed population



Source: Prognos AG/DevStat (2021).

Besides these regional differences, there are also some limitations concerning LinkedIn's sectoral and educational representativeness.⁶¹ In addition to the analysis of user groups in Table 2-1, it needs to be highlighted that some industries (e.g., financial services) and especially white-collar workers are overrepresented on LinkedIn. Similarly, high educated individuals with at least a bachelor's degree and/or people who want to network are more likely to use LinkedIn compared to the population with lower levels of educational attainments. A further limitation could be the not clearly defined technologies and the relevant skills. The threshold area can therefore vary depending on the definition of the technology. Additionally, the information in each LinkedIn profile is self-reported. i.e., there is no guarantee that the LinkedIn members have the specified skills or live in the specified region. Therefore, "LinkedIn Members with specific skills" in the following refers to the self-reported skills of each user that are queried with the keywords.

5.2.2 Illustrations of findings on technical skills for digital innovation

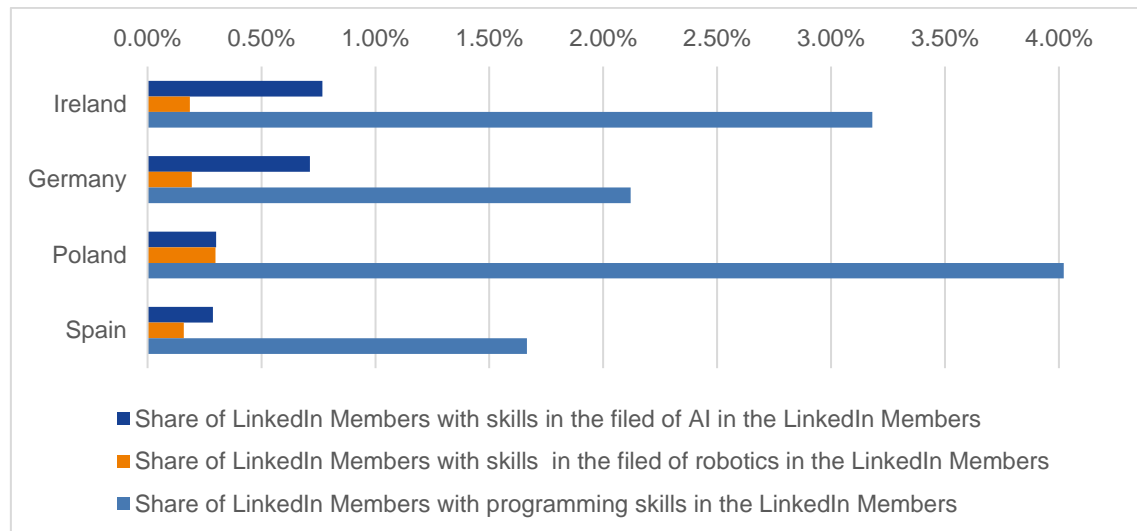
Comparing the programming skills based on LinkedIn on a national level, shown in Figure 5-2, Poland has the highest share of LinkedIn members with programming skills specify in their profile in the total LinkedIn members in this country (~4.00%), followed by Ireland (~3.18%), Germany (~2.12%) and Spain (1.76%). If one relates the share of LinkedIn members with programming skills to the active population of the country, shown in Figure 5-3, the figures shift. The active population (labour force) is defined as the sum of employed and unemployed persons and is equivalent to the expression 'person in the labour force'.⁶² Related to the active population, Ireland has the largest share of LinkedIn members with programming skills at 2.65%.

⁶¹ World Bank Group/LinkedIn (n.d.): Data Insights: Jobs, Skills and Migration Trends Methodology & Validation Results. <http://documents1.worldbank.org/curated/en/827991542143093021/pdf/World-Bank-Group-LinkedIn-Data-Insights-Jobs-Skills-and-Migration-Trends-Methodology-and-Validation-Results.pdf> and European Commission (2020): Advanced Technologies for Industry - Methodological report. <https://ati.ec.europa.eu/reports/eu-reports/advanced-technologies-in-industry-methodological-report>

⁶² https://ec.europa.eu/eurostat/cache/metadata/en/lfsi_esms.htm

Spain, Poland, and Germany follow with 1.05%, 0.92%, and 0.61% respectively. It follows that Poland does have a relatively high proportion of programmers among its LinkedIn members. However, in relation to the active population Ireland has the highest share. The background of this difference is the high share of LinkedIn members in Ireland of around 38% in the total and 83% in the employed population compared to the low proportion in Poland of around 10% in the total and 23% in the employed population. On a national level, Ireland has the highest share of LinkedIn Members who specify AI skills in their profile in the total LinkedIn members (0.77%), followed by Germany (0.71%), Poland (0.30%), and Spain (0.29%). Robotics, on the other hand, has the highest share in Poland at 0.30% followed by Ireland and Germany at 0.19% and Spain at 0.16%. For comparison of the figures with official statistics, the Eurostat database⁶³ has been used.

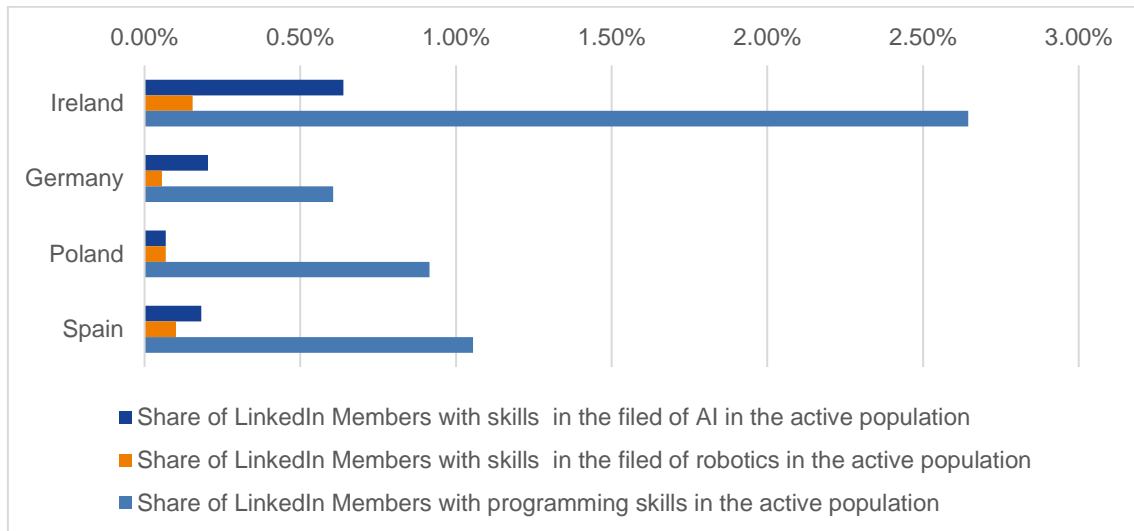
Figure 5-2: Share of LinkedIn Members with skills in the field of Programming, Robotics, and AI in the total LinkedIn Members by country



Source: Prognos AG/DevStat (2021).

⁶³ <https://ec.europa.eu/eurostat/de/data/database>

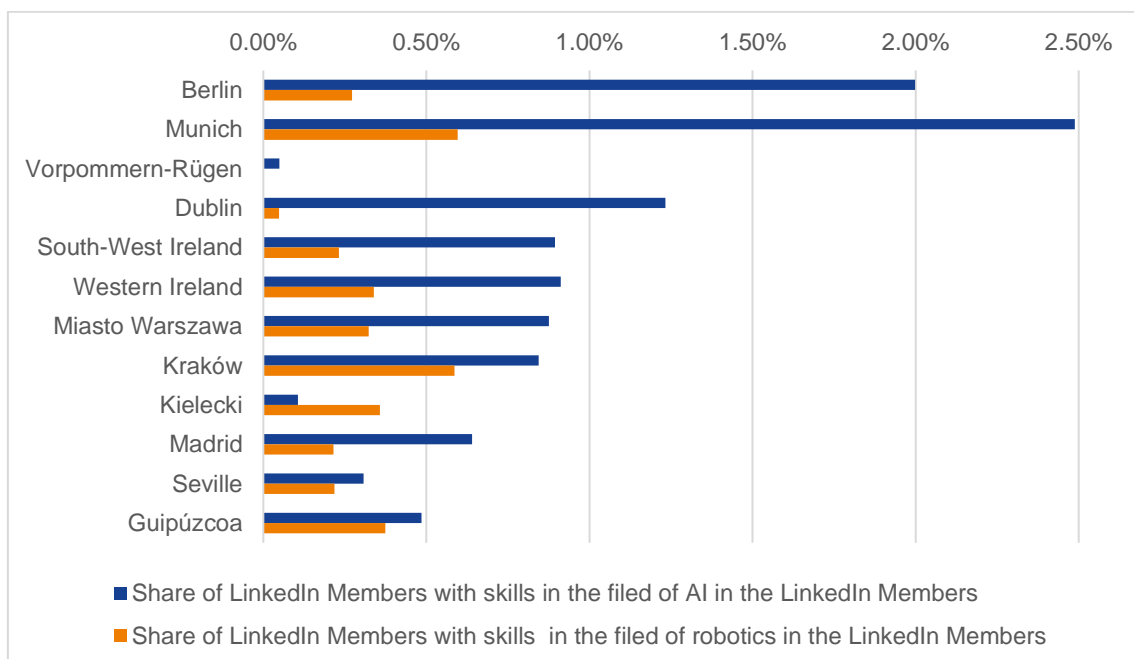
Figure 5-3: Share of LinkedIn Members with skills in the field of Programming, Robotics, and AI in the active population by country



Source: Prognos AG/DevStat (2021).

Looking at the share of LinkedIn members with skills in the field of AI in the total LinkedIn Members on a regional level, Munich has the highest share with almost 2.5% followed by Berlin (2.0%) and Dublin (1.23%). In the field of robotics, Munich and Kraków show the highest shares of LinkedIn Members (~0.6%). Across almost all regions, the share of LinkedIn members with skills in AI is higher than in Robotics. This could be since the keywords were rather narrowly defined (only robotics, no keywords on automation, etc.), whereas the keywords in the field of AI cover many dimensions like fields of application or AI methods. The only region with a significantly higher share of LinkedIn members in Robotics is Kielecki (Poland). No LinkedIn member stated to live in the region Vorpommern-Rügen and to have skills in the field of robotics.

Figure 5-4: Share of LinkedIn Members with skills in the field of AI and Robotics in the total LinkedIn Members by region

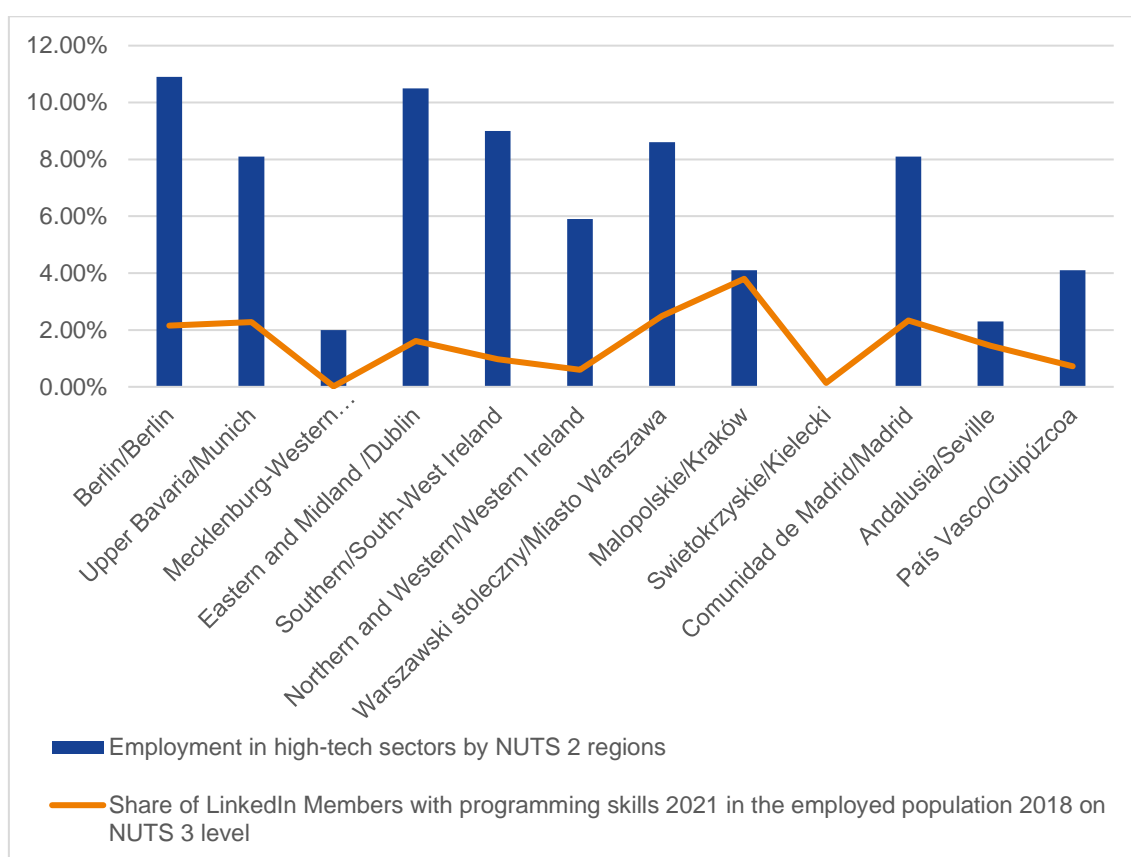


Source: Prognos AG/DevStat (2021).

Comparison with official statistics

To assess whether LinkedIn is suitable as a data source for official statistics, the data is compared with existing official statistics. For this purpose, data of 'Employment in high-tech sectors' from the Eurostat database⁶⁴ has been used. Data on LinkedIn Members with programming skills 2021 in the employed population 2018 is compared with official statistics on high-tech industry and knowledge-intensive services 2019. Since 2019 data were not available for the two regions Mecklenburg-Western Pomerania and Swietokrzyskie, 2018 data were used for these two regions. An overview of how *High-tech industry and knowledge-intensive services* can be found in the Explanatory texts in Annex 2 – Annex 6 in that document.⁶⁵ It needs to be highlighted that the Eurostat data is only available on the NUTS 2 level and therefore the comparability is limited. Nonetheless, in absence of more granular data, this Eurostat database is used for an approximate approach of comparing the data and some conclusions can be drawn.

Figure 5-5: Compare the share of LinkedIn Members with programming skills in the employed population on NUTS 3 level with employment in high-tech sectors by NUTS 2 region



Source: Prognos AG/DevStat (2021).

Overall, a rather higher rate of LinkedIn Members with programming skills is more likely to be found in regions with a higher employment rate in high-tech sectors. Munich, Seville or Krakow are exceptions. When interpreting these results, the different limitations must be considered. A limitation already mentioned are the different NUTS levels, which lead to the fact that in Berlin, for example, an almost identical area (Berlin

⁶⁴ <https://ec.europa.eu/eurostat/de/data/database>

⁶⁵ https://ec.europa.eu/eurostat/cache/metadata/Annexes/htec_esms_an2.pdf

at NUTS 2 and NUTS 3 level) is compared, while in Munich and Bavaria a comparison is made between a large city and an entire state. In addition, data from different years are compared. In a field as dynamic as technology, a lot can change in just one year. To address these two limitations, it is useful to look at time series.

Despite the different NUTS levels, the analysis shows that LinkedIn can provide relevant data for official statistics which could potentially provide comprehensive insights into topics relating to employment, economic development, digitalization, and further. LinkedIn data offers added value in particular through a finer regional granularity (NUTS 3 level) and more differentiated topic analyses (e.g., skills in the fields of AI or programming skills as different granular subcategories of digital skills). If time series are available, insights into developments and trends can be gained.

5.3 Twitter: COVID-19

For the assessment of topics around COVID-19 and sustainable mobility, Twitter was chosen as a relevant data source. Two types of indicators are assessed in the selected pilot regions:

- TI1_12: Number of tweets on the topic covid
- TI1_13: Number of tweets on the topic covid per 10000 capita
- TI1_14: Number of tweets on vaccines from capital cities per 10,000 capita

Like the descriptions for LinkedIn, we first describe the methodology of how to access the data, followed by illustrative findings from this analysis, separated into findings on COVID-19. We close the section with some further options for analysis, that were identified during the work on the demonstrator study.

5.3.1 Methodology

To access Twitter data, the provided standard API from Twitter has been used. Therefore, it is required to apply and receive approval for a developer account. Applying for the standard API is the easiest way to get access to Twitter data, as the approval process is relatively quick and straightforward. The default product track is well suited for demonstration and learning but has its limitations with respect to time-series data since only tweets from the last 7 days can be queried by keyword.⁶⁶ The query for this demonstrator study was conducted on 02.09.21 and accordingly contains tweets from 24 August to 02 September 2021.

An ontology was created to retrieve corresponding tweets on the topics Covid 19 (Table 5-3). The ontology for the topic Covid 19 is based on *COVID-19 Vaccination Awareness and Aftermath*⁶⁷ and was enriched with further terms. The ontologies were translated into all relevant languages (German, Polish, and Spanish). The ontology was used to query the tweets systematically and automatically on the topic named topics. Longitude and latitude coordinates and a radius were used for the retrieval of tweets in each region. Through this process, the boundaries of the selected NUTS 3 level regions are set approximately and hence it could be the case that the tweets of a region containing a few tweets, which are geotagged with a neighbouring region. An OR query was executed per label and region, i.e., tweets located in the selected regions containing at least one of the keywords per label were queried.⁶⁸ Further query options can be found on the Twitter developer platform.⁶⁹ Subsequently, the returned data was cleaned. For example, the tweets from bots were deleted and duplicates, which arose due to language overlaps (for example COVID19 is the same word in every language), were removed.

⁶⁶ <https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api>

⁶⁷ Sattar, N.S.; Arifuzzaman, S. COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA. Appl. Sci. 2021, 11, 6128.

⁶⁸ <https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query>

⁶⁹ <https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream/integrate/build-a-rule>

Table 5-3: Ontology on the Topic Covid 19

Label	Keywords
Pfizer	pfizer; Comirnaty; Pfizer-BioNTech; BioNTechpfizer
Moderna	Moderna; moderna_tx; Moderna-NIAID; NIAID; NIAID-Moderna
Johnson & Johnson	Johnson & Johnson; Johnson and Johnson; Janssen; Janssen Pharmaceutical; J&J
Oxford-AstraZeneca	OXFORDVACCINE; Oxford-Astraeneca; OxfordAstraZeneca; AstraZeneca; Vaxzevria; Covishield
SputnikV	Sputnik V; sputnikv; sputnikvaccine
Covaxin	covaxin; BharatBiotech
Sinovac	coronavac; sinovac
Types of vaccines	RNA vaccine; mRNA vaccine; vector vaccine
Hygiene	hand sanitizer; sanitizer; wash hands; wash face; soap; soap water; hand soap
Wear mask	sanitize mask; wearamask; masking; N95; face cover; face covering; face covered; mouth cover; mouth covering; mouth covered; nose cover; nose covering; nose covered; cover your face; coveryourface
Travel	travel; outing; camping; air-travel
Social Distancing	social distancing; physical distancing; 6 feet; social distance; physical distance; social gathering; gathering; party; restaurant;
Social Gathering	social gathering; gathering; party; restaurant
Vaccination scepticism	adverse effects; side effects; antivax; immune; herd immunity; vaccine scepticism; vaccine refusal; vaccine hesitancy; anti-vaccination
Corona	COVID19; COVID-19; coronavirus; corona; pandemic

Source: Prognos AG/DevStat (2021).

In addition to the analysis of user groups as outlined in Table 2-1 further characteristics of Twitter users can be identified. The first concerns the Twitter penetration rate (i.e., the number of Twitter users compared to the total population) across the different European countries and regions.⁷⁰ Here, the share of Twitter users is significantly higher in countries like Spain, the UK and France compared to Germany and many eastern European regions. Another factor is the age of Twitter users as some studies find that the most geotagged tweets are emitted by users that are between 15 and 30 years old.⁷¹ Moreover, some studies suggest that the reasons for posting geo-tagged tweets are connected to a higher social-economic status or education of users.⁷² As a concluding remark, it must be mentioned that Twitter users can have several accounts. This means that Tweets from different user IDs in the dataset could have been posted by the same person.⁷³

⁷⁰ Lenormand et al. (2014): Tweets on the road. Instituto de Fisica Interdisciplinar y Sistemas Complejos

⁷¹ Sloan, L., & Morgan, J. (2015): Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. PloS one, 10(11), e0142209. <https://doi.org/10.1371/journal.pone.0142209>

⁷² Martí et al. (2019): Social Media data: Challenges, opportunities and limitations in urban studies In Computers, Environment and Urban Systems, Vol 74, p. 161-174

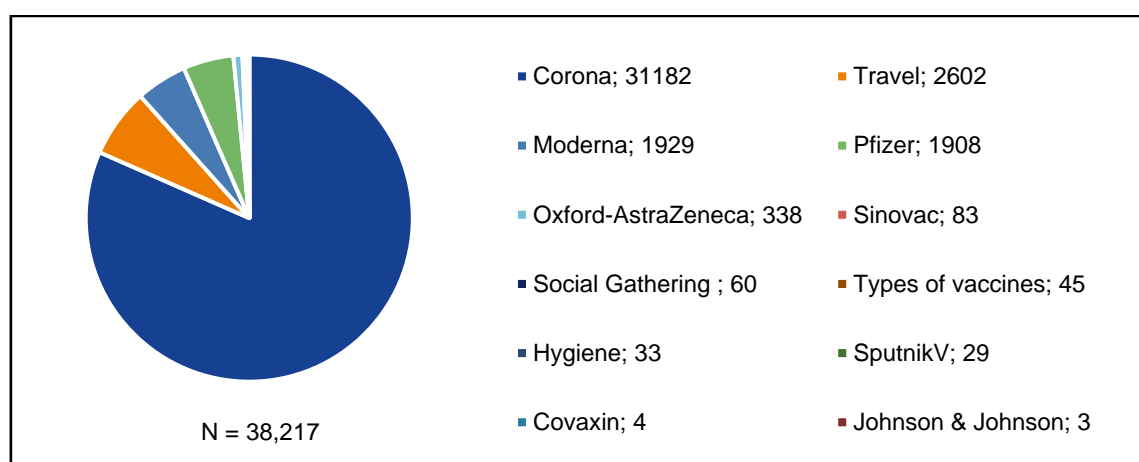
⁷³ see also European Commission (2019): Measuring labour mobility and migration using big data. Available online: <https://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=8264>

5.3.2 Illustrations of findings on Covid-19

In total, 38.217 tweets and 37.300 distinct tweet IDs were returned. The gap is caused by the multiple assignments of tweets to different labels. For example, a tweet containing BioNTechpfeizer and coronavirus is assigned to both Pfizer and Corona labels and therefore appears twice in the list with the same tweet ID.

Across regions, one can state that people talk most about Corona in general as shown in Figure 5-6. Analysing individual vaccines separately, the mRNA vaccines from Pfizer and Moderna are mentioned significantly more frequently in the underlying tweets than other vaccines such as Astra Zeneca or Sinovac. The number of Tweets with the keywords on the topics of social gathering, types of vaccines and hygiene, and vaccination scepticism is also comparatively low. A graphical overview of the exact numbers of tweets per label is given in Figure 5-6.

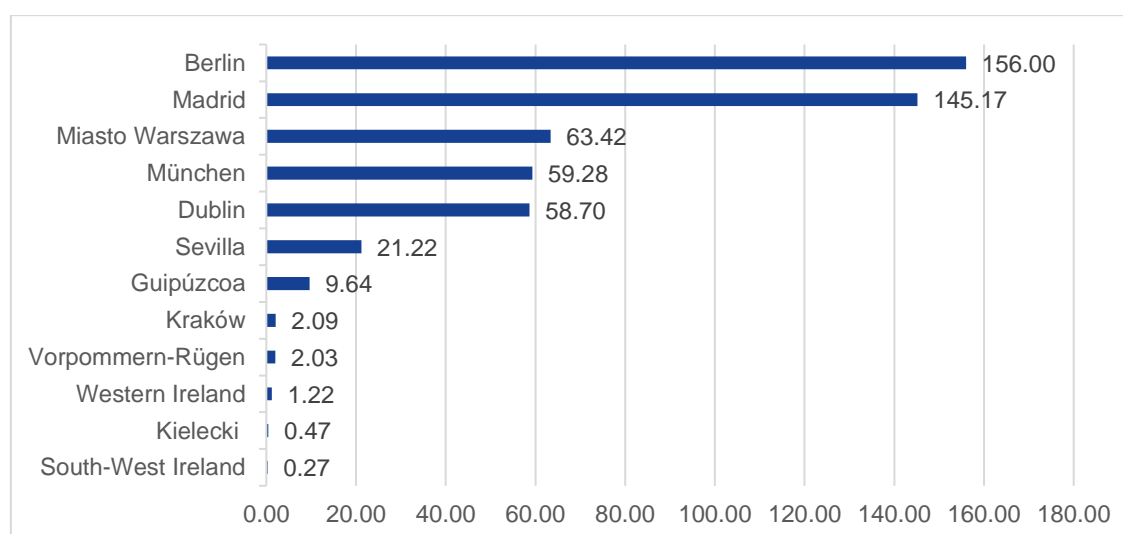
Figure 5-6: Number of tweets on Covid-19 per label (subtopic)



Source: Prognos AG/DevStat (2021).

Overall, the number of tweets about Corona tends to be higher in urban areas and metropolitan regions. Figure 5-7 shows the number of tweets based on the population size of the region. In this context, Berlin (156) and Madrid (~145) have the highest number of tweets. The two regions Kielecki and South-West Ireland have the fewest tweets based on population size.

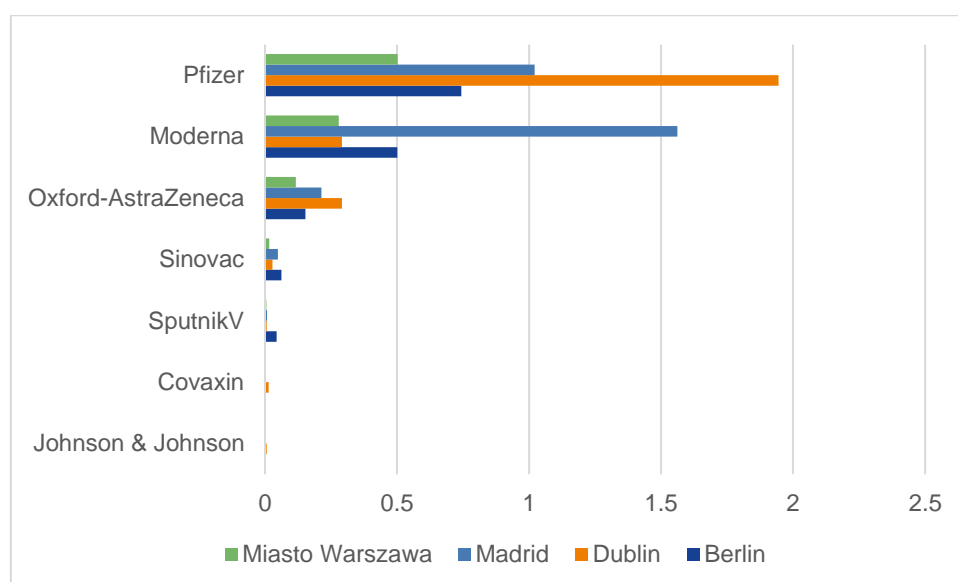
Figure 5-7: Number of tweets on Covid-19 per 10,000 inhabitants



Source: Prognos AG/DevStat (2021).

To get an overview of tweets about different vaccines, the number of tweets by vaccine per 10000 population is mapped in Figure 5-8. Comparing the different vaccines, it shows that overall, it is most tweeted about the Pfizer vaccine, followed by Moderna and Oxford-AstraZeneca. While in Warsaw, Dublin, and Berlin more tweets per inhabitant were about the Pfizer vaccine than about Moderna, the tweets in Madrid contained more terms about Moderna than Pfizer. The Covaxin and Johnson & Johnson vaccines are very rarely mentioned in the tweets analysed. The analysis shows how different topics (in this case vaccines) in different regions can be presented and analysed based on twitter. The indicator shows, by way of example, that different regions tweet about vaccines with varying frequency over a given period of time.

Figure 5-8: Number of tweets on vaccines from capital cities per 10,000 inhabitants



Source: Prognos AG/DevStat (2021).

The analysis and presentation of the tweets on the topic of Corona exemplifies the possibilities opened by a systemic analysis of tweets. Such an evaluation also poses challenges. Tweets must be cleaned of bots, for example. Technically, the geolocation of tweets poses a particular challenge. In the demonstration study, regions were only approximated using longitude and latitude. For a large-scale analysis of the different regions in the EU, it makes sense to further refine this approach to identify the regions more precisely. When analysing and interpreting the results, it is important to consider that only a subset of the tweets is geotagged. In addition, the effort required to translate the ontology into different languages and loops to improve the ontology for quality assurance and testing must always be considered. An extended API, such as Twitter's Premium API, can help address some of the challenges mentioned above and provide additional opportunities such as time series analysis.

Comparison with official statistics

Big data territorial indicators are computed at high geographical disaggregation levels (NUTS3 level in the case of the proposed COVID19 indicators), which is in general beyond the breakdowns provided by official data. However, it is possible to make a comparison between the COVID19 indicators computed in this demonstration study with the information on the number of infections and vaccination levels provided by the European Centre for Disease Prevention and Control (ECDC) at the country level, as presented in Table 3-5.

Table 5-4: Infections, vaccination, and salience of COVID19

NUTS0 area	NUTS3 area	14-day case notification rate per 100 000 inhabitants at NUTS0 level (1)	% full vaccinated population at NUTS0 level (2)	Number of tweets in relation to population at NUTS3 level. (Factor 10,000)
Germany	Berlin	137,0	60,3	31,9
Germany	Munich	137,0	60,3	28,6
Germany	Vorpommern-Rügen	137,0	60,3	13,4
Ireland	Dublin	504,1	69,4	32,3
Ireland	Galway (Western)	504,1	69,4	17,6
Ireland	Cork (South-West)	504,1	69,4	1,4
Poland	Miasto Warszawa	7,7	48,0	8,0
Poland	Kraków	7,7	48,0	5,3
Poland	Kielecki	7,7	48,0	0,9
Spain	Madrid	270,5	66,9	15,3
Spain	Seville	270,5	66,9	19,4
Spain	Guipúzcoa	270,5	66,9	6,3

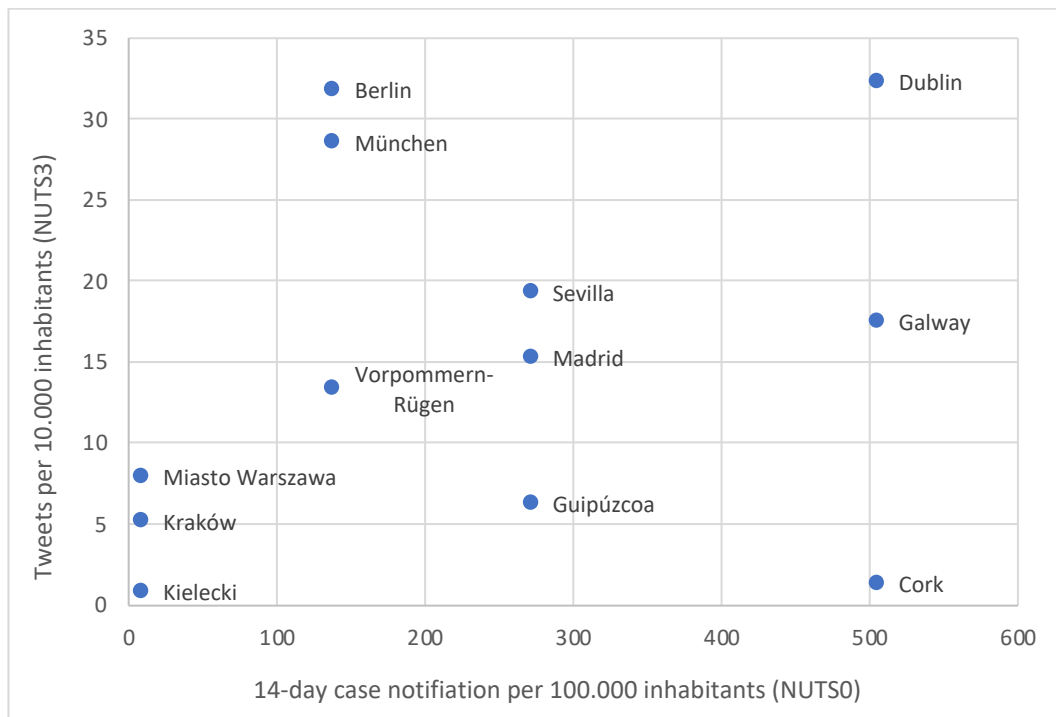
(1) Notifications from 16 to 29 August 2021 at www.ecdc.europa.eu/en/cases-2019-ncov-eueee

(2) Vaccination rates on 7 September 2021 at <https://vaccinetracker.ecdc.europa.eu/public/extensions/COVID-19/vaccine-tracker.html#uptake-tab>

Source: Prognos AG/DevStat (2021). Own elaboration from European Centre for Disease Prevention and Control data

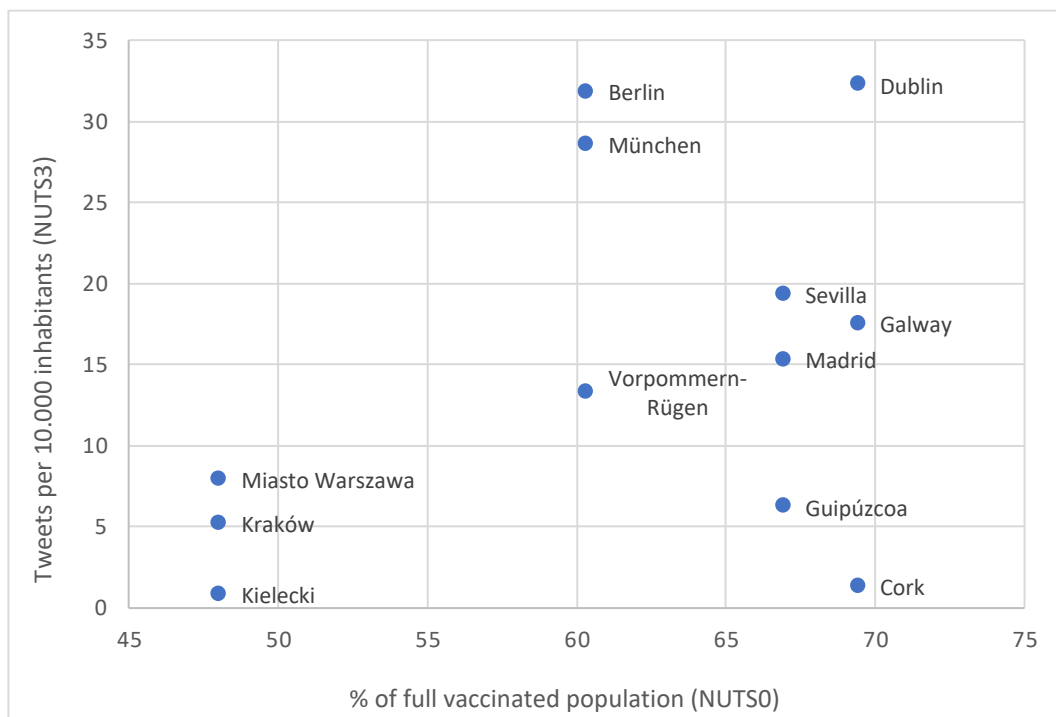
As shown in Figure 5-9 it seems to be a direct relation between the number of infections in the country and the salience of COVID19 in most of the NUTS3 areas, measured as the number of tweets referring to the pandemic in each NUTS 3 area (with the exceptions of Cork in Ireland and Guipúzcoa in Spain). This fact suggests that, beyond the straightforward information provided by territorial big data indicators, they can be used as independent variables in a small area and nowcasting models of other variables provided by official statistics for larger areas and with lower frequency. A similar relation between vaccination levels and salience of the disease is presented in Figure 5-10.

Figure 5-9: Number of infections and COVID19 salience



Source: Prognos AG/DevStat (2021).

Figure 5-10: Vaccination rates and COVID19 salience



Source: Prognos AG/DevStat (2021).

6 Concluding remarks and outlook

Territorial indicators based on big data offer a multitude of new possibilities. For policy making, economic decision makers, and society, the knowledge from the spatial patterns of Big Data offers important insights and the basis for smart action. In addition, the intelligent use and analysis of this amount of data is increasingly becoming a critical competitive parameter for companies and authorities alike.

This experimental study explored various possibilities for the analysis of data from private, digital platforms. A key result is a list of private, digital platforms that are particularly suitable for such analysis. In a multi-stage process, these platforms were derived from a long list of almost 80 platforms. Several different perspectives were used for this assessment. These include, for instance, criteria such as a Europe-wide coverage of these platforms, data that is as comparable as possible in all European regions, the (technical and legal) accessibility of the data and, in how far the data provides insights into topics of the Territorial Agenda 2030. A central criterion in all this is the location of the information. Digital platforms have no place in a classic sense. The platforms are located in a virtual room. The location of the server cannot be directly experienced by the user. Only indirectly, for example through the legal framework to which the platforms are subject. Nevertheless, there are numerous 'territorial traces' left by the users of the platforms. These territorial traces are used for the subsequent processing of the information from the platforms in spatial analyses.

An important result of this study refers to the large differences between the private digital platforms. Because of these differences, there is not just one way to obtain spatial information from these platforms. Some of the possible approaches have been presented in the analysis of the platforms and the demonstrator study. Overall, data from private digital platforms creates many new possibilities:

- **New thematic possibilities:** The wealth of information offered by private digital platforms is also reflected in completely new possibilities to store abstract constructs such as quality of life using indicators. The assessment of texts or satellite images or network structures, for example, significantly expands the spectrum of information acquisition. It has been shown in this study, that the subject areas of the Territorial Agenda 2030 can be measured using these new indicators.
- **New methodological possibilities:** The information that can be obtained from the digital platforms only has a short collection time compared to the data basis of classic indicators. Some of these are available almost in real-time. The possibility to collect populations and not only samples is usually not given with classic indicators. The mass of information also offers the opportunity to differentiate the information in a variety of ways. In this context, the data must not be understood as a sample of the entire population. Rather, the goal should be to make particular use of the special information content, the strengths of this data. New methodological possibilities also arise through the use of these indicators, e.g., nowcasting, the calculation of regional trends, and a variety of comparison options.

The variety of possibilities was also the centre of the discussion by two focus groups conducted as part of this study. In this group of experts, challenges in accessing data on private digital platforms were discussed. It became clear, that data access through rather complex legal agreements is not the primary solution but rather a flexible approach utilizing the existing data access options should be explored. Moreover, the experts helped to identify subject areas which are particularly suitable for indicators from data of private digital platforms.

After exploring the full range of possibilities, we propose to focus more on certain types of territorial indicators in future studies. This allows the findings of this study to be deepened and supplemented. An analysis of territorial timeseries data and nowcasting is especially promising. This offers the possibility to assess the volatility and seasonality. In addition, nowcasting is interesting because it represents a connection between previous classic indicators and the new big data indicators. The periods in which both indicators are available are used to estimate the extent to which the classic indicators can be predicted by the new indicators. This also offers the possibility of better assessing the validity of the results in addition to the reliability of the data.

All in all, this study helped to explore a new methodological territory. Particularly promising platforms were selected, and a list of possible territorial indicators was presented. The legal framework in which this data can be used has been explained - a path that is essentially based on the terms of use of the platform operator. Finally, different territorial indicators were collected and evaluated as part of a small demonstrator study, which looked more at the feasibility of generating the data for a selection of platforms and indicators rather than performing a comprehensive (comparative) analysis as often the case in ESPON projects.

7 Annex

7.1 Long-list of selected platforms

Table 7-1: Full long list of 79 digital platforms providing big data

Platforms	Thematic field	Type of service	Territorial Traces / Territorial data traces	Typology (type of resource granted access to)	Data scope
Flickr	Picture Sharing	Social media platforms	territorial traces via user-registration (geotagging etc.)	Access to personal data and other private content	EU-27 + EFTA + UK
Instagram	Picture Sharing	Social media platforms	territorial traces via user-registration (geotagging etc.)	Access to personal data and other private content	EU-27 + EFTA + UK
Facebook	News/Information/Picture Sharing	Social media platforms	territorial traces via user-registration (geotagging etc.)	Access to personal data and other private content	EU-27 + EFTA + UK
Twitter	News/Information/Picture Sharing	Social media platforms	territorial traces via user-registration (geotagging etc.)	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
LinkedIn	Employment	Social media platforms	territorial traces via user-registration (geotagging etc.)	Access to personal data and other private content	EU-27 + EFTA + UK
Google Maps	Rout planning/ Location identification	Web mapping service	direct territorial traces available	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Foursquare	Tourism platform	Recommendations platforms	territorial traces via user-registration (geotagging etc.)	Access to goods and services offered by third parties (online markets or sharing economy platforms)	EU-27 + EFTA + UK
TripAdvisor	Tourism platform	Recommendations platforms	Web-portal/Website with territorial Information	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Tumblr	News/Information/Picture Sharing	Social media platforms	territorial traces via user-registration (geotagging etc.)	Access to personal data and other private content	EU-27 + EFTA + UK
Booking	Tourism platform	Online travel agency	Web-portal/Website with territorial Information	Access to goods and services offered by third parties (online markets or sharing economy platforms)	EU-27 + EFTA + UK
Microsoft Academic	Knowledge platforms	Meta search engine	territorial traces via user-registration (geotagging etc.)	Access to goods and services offered by third parties (online markets or sharing economy platforms)	EU-27 + EFTA + UK

Platforms	Thematic field	Type of service	Territorial Traces / Territorial data traces	Typology (type of resource granted access to)	Data scope
ResearchGate	Knowledge and Social media platforms	Social media platforms	territorial traces via user-registration (geotagging etc.)	Access to goods and services offered by third parties (online markets or sharing economy platforms)	EU-27 + EFTA + UK
Academia	Knowledge and Social media platforms	Social media platforms	territorial traces via user-registration (geotagging etc.)	Access to goods and services offered by third parties (online markets or sharing economy platforms)	EU-27 + EFTA + UK
Google Scholar	Knowledge platforms	Meta search engine	Web-portal/Website with territorial Information	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Airbnb	Lodging	Marketplace	Web-portal/Website with territorial Information	Access to goods and services offered by third parties (online markets or sharing economy platforms)	EU-27 + EFTA + UK
skyscanner	Mobility	Meta search engine	Web-portal/Website with territorial Information	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Kayak	Mobility	Meta search engine	Web-portal/Website with territorial Information	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Expedia	Tourism platform	Online travel agency	Web-portal/Website with territorial Information	Access to goods and services offered by third parties (online markets or sharing economy platforms)	EU-27 + EFTA + UK
Trivago	Tourism platform	Online travel agency	Web-portal/Website with territorial Information	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Strava	Sports	Social media platforms	direct territorial traces available	Access to goods and services offered by third parties (online markets or sharing economy platforms)	EU-27 + EFTA + UK
Pinterest	Picture/Ideas Sharing	Social media platforms	territorial traces via user-registration (geotagging etc.)	Access to personal data and other private content	EU-27 + EFTA + UK
TikTok	Video Sharing	Social media platforms	territorial traces via user-registration (geotagging etc.)	Access to personal data and other private content	EU-27 + EFTA + UK
Indeed	Employment	Meta search engine	Web-portal/Website with territorial Information	Access to goods and services offered by third parties (online markets or sharing economy platforms)	Available in 19 EU countries
monster	Employment	Meta search engine	Web-portal/Website with territorial Information	Access to goods and services offered by third parties (online markets or sharing economy platforms)	Available in 19 EU countries
BlaBlaCar	Mobility	Marketplace	territorial traces via user-registration (geotagging etc.)	Access to goods and services offered by third parties (online markets or sharing economy platforms)	available in 14 EU MS
Stepstone	Employment	Meta search engine	Web-portal/Website with territorial Information	Access to goods and services offered by third parties (online markets or sharing economy platforms)	available in 24 EU MS

Platforms	Thematic field	Type of service	Territorial Traces / Territorial data traces	Typology (type of resource granted access to)	Data scope
Flixbus	Mobility	Intercity bus service	Web-portal/Website with territorial Information	No assignment possible	Excluding EL, MT, FI, LT, LV, EE, CY
Yelp	Business reviews	Recommendations platforms	Web-portal/Website with territorial Information	Access to information such as general search engines or specialised search engines	only operating in a few EU MS
Uber	Mobility	Marketplace	territorial traces via user-registration (geotagging etc.)	Access to goods and services offered by third parties (online markets or sharing economy platforms)	only operating in 17 EU MS, only selected cities
Deliveroo	Food	Online delivery service	Web-portal/Website with territorial Information	Access to goods and services offered by third parties (online markets or sharing economy platforms)	only operating in NL, FR, BE, IE, ES, IT
Delivery Hero	Food	Online delivery service	Web-portal/Website with territorial Information	Access to goods and services offered by third parties (online markets or sharing economy platforms)	only operating in a few EU MS
Foodpanda	Food	Online delivery service	Web-portal/Website with territorial Information	Access to goods and services offered by third parties (online markets or sharing economy platforms)	only operating in RO, BG
Bolt	Mobility	Marketplace	Web-portal/Website with territorial Information	Access to goods and services offered by third parties (online markets or sharing economy platforms)	only operating in a few EU MS
Etsy	Goods	Marketplace	indirect (no locational information without IP)	Access to goods and services offered by third parties (online markets or sharing economy platforms)	EU-27 + EFTA + UK
Youtube	Video sharing	Social media platforms	indirect (no locational information without IP)	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Stackoverflow	Programming	Internet forum	indirect (no locational information without IP)	Access to goods and services offered by third parties (online markets or sharing economy platforms)	EU-27 + EFTA + UK
Github	Programming	Code hosting platform	indirect (no locational information without IP)	Access to goods and services offered by third parties (online markets or sharing economy platforms)	EU-27 + EFTA + UK
Paypal	Finance	Payment service	territorial traces via user-registration (geotagging etc.)	Access to money or capital or payment systems	EU-27 + EFTA + UK
Quora	Question/answer	Internet forum	indirect (no locational information without IP)	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Stackexchange	Programming	Internet forum	indirect (no locational information without IP)	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Udemy	Programming	Online learning platform	indirect (no locational information without IP)	Access to goods and services offered by third parties (online markets or sharing economy platforms)	EU-27 + EFTA + UK

Platforms	Thematic field	Type of service	Territorial Traces / Territorial data traces	Typology (type of resource granted access to)	Data scope
SlideShare	Document sharing	Social media platforms	indirect (no locational information without IP)	Access to personal data and other private content	EU-27 + EFTA + UK
Vimeo	Video sharing	Media Sharing	indirect (no locational information without IP)	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Reddit	Discussion	Social-News-Aggregator	indirect (no locational information without IP)	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Spotify	Music	Streaming service	indirect (no locational information without IP)	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Netflix	Movies	Streaming service	indirect (no locational information without IP)	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Deezer	Music	Streaming service	indirect (no locational information without IP)	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Apple Music	Music	Streaming service	indirect (no locational information without IP)	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Geocaching	Leisure	Web mapping service	indirect (no locational information without IP)	Access to information such as general search engines or specialised search engines	EU-27 + EFTA + UK
Ebay	Goods	Marketplace	indirect (no locational information without IP)	Access to goods and services offered by third parties (online markets or sharing economy platforms)	only available in 10 EU MS
Amazon	Goods	Marketplace	indirect (no locational information without IP)	Access to goods and services offered by third parties (online markets or sharing economy platforms)	not available in entire EU, but: no amazon.be, but Belgium users can order from amazon.de
Zalando	Clothing	Marketplace	indirect (no locational information without IP)	Access to goods and services offered by third parties (online markets or sharing economy platforms)	only available in 15 EU MS
Kickstarter	Crowdfunding	Crowdfunding	indirect (no locational information without IP)	Access to money or capital or payment systems	only operating in a few EU MS
GoFundMe	Crowdfunding	Crowdfunding	indirect (no locational information without IP)	Access to money or capital or payment systems	only operating in 13 EU MS
Bitcoin	Finance	Cryptocurrency	indirect (no locational information without IP)	Access to money or capital or payment systems	EU-27 + EFTA + UK
Weather	Weather	News	Web-portal/Website with territorial Information	No assignment possible	EU-27 + EFTA + UK

Platforms	Thematic field	Type of service	Territorial Traces / Territorial data traces	Typology (type of resource granted access to)	Data scope
Amadeus	Tourism	IT-Service	Web-portal/Website with territorial Information	No assignment possible	EU-27 + EFTA + UK
Travelport	Tourism	IT-Service	Web-portal/Website with territorial Information	No assignment possible	EU-27 + EFTA + UK
Europcar	Mobility	Online car rental service	Web-portal/Website with territorial Information	No assignment possible	EU-27 + EFTA + UK
Hertz	Mobility	Online car rental service	Web-portal/Website with territorial Information	No assignment possible	EU-27 + EFTA + UK
Avis	Mobility	Online car rental service	Web-portal/Website with territorial Information	No assignment possible	EU-27 + EFTA + UK
Sixt	Mobility	Online car rental service	Web-portal/Website with territorial Information	No assignment possible	EU-27 + EFTA + UK
Marine Traffic	Mobility	Live Ships Map	direct territorial traces available	No assignment possible	EU-27 + EFTA + UK
TomTom	Mobility	Provider of geodata	direct territorial traces available	No assignment possible	EU-27 + EFTA + UK (BG / CY only partially covered)
HERE	Mobility	Provider of geodata	direct territorial traces available	No assignment possible	EU-27 + EFTA + UK
Garmin	Sports	Provider of geodata	direct territorial traces available	No assignment possible	EU-27 + EFTA + UK
Flightradar24	Mobility	Live Flight Map	direct territorial traces available	No assignment possible	EU-27 + EFTA + UK
UPS (United Parcel Service)	Logistics	Package delivery	Web-portal/Website with territorial Information	No assignment possible	Package operations EU-wide + EFTA + UK
IATA (International Air Transport Association)	Aircraft	Industry association	direct territorial traces available	No assignment possible	EU-27 + EFTA + UK
Komoot	Leisure	Web mapping service	direct territorial traces available	No assignment possible	only available in 15 EU MS + Switzerland

Platforms	Thematic field	Type of service	Territorial Traces / Territorial data traces	Typology (type of resource granted access to)	Data scope
Hermes Europe	Logistics	Package delivery	Web-portal/Website with territorial Information	No assignment possible	only available in 24 countries in the EU, some only with partners
Visa	Finance	Credit card service	indirect (no locational information without IP)	Access to money or capital or payment systems	EU-27 + EFTA + UK
Master Card	Finance	Credit card service	indirect (no locational information without IP)	Access to money or capital or payment systems	EU-27 + EFTA + UK
DHL	Logistics	Package delivery	indirect (no locational information without IP)	No assignment possible	EU-27 + EFTA + UK
Wordpress	Website-building	Content management system	indirect (no locational information without IP)	No assignment possible	EU-27 + EFTA + UK
Shop Apotheke Europa	Pharmacy	Web shop	indirect (no locational information without IP)	No assignment possible	only available in FR, BE, NL, DE, AU
Shopify	Website-building	E-Commerce-Software	indirect (no locational information without IP)	No assignment possible	Available in 8 EU countries
o2	ICT	Mobile communications	indirect (no locational information without IP)	No assignment possible	no coverage of many ESPON countries
Vodafone	ICT	Mobile communications	indirect (no locational information without IP)	No assignment possible	no coverage of many ESPON countries

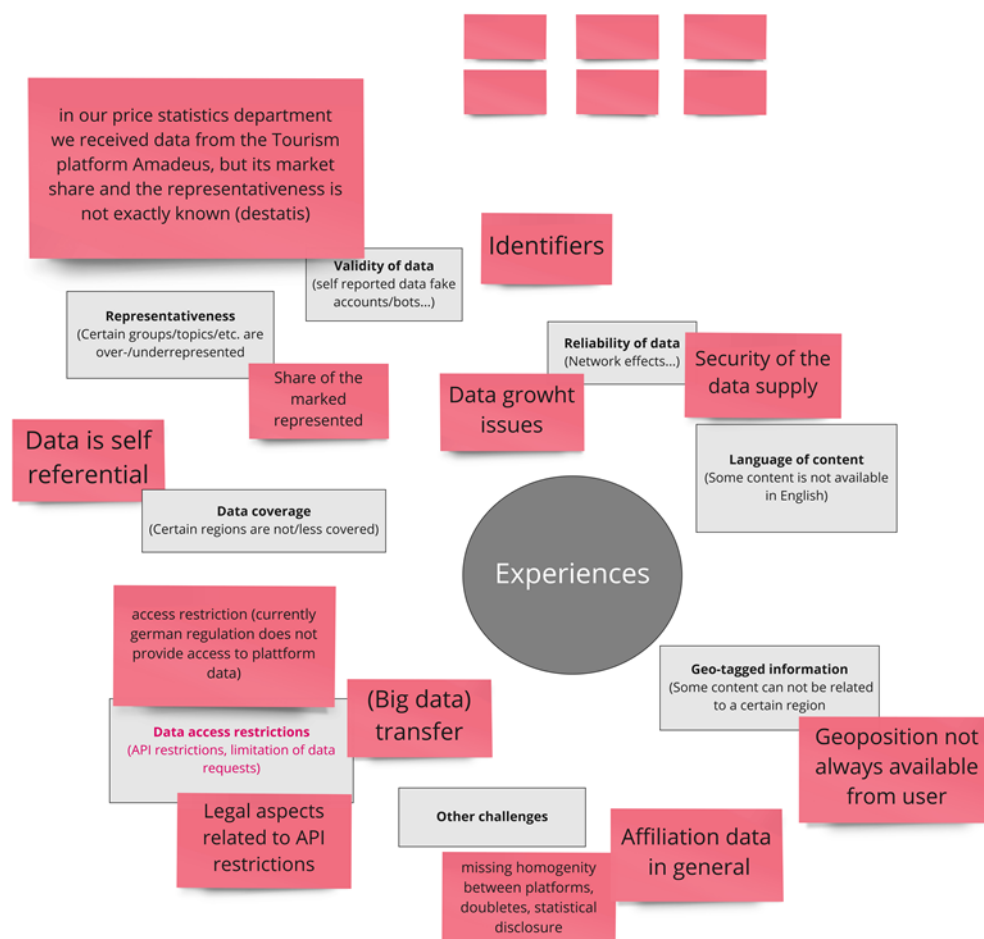
Source: Prognos AG/DevStat (2021).

7.2 Results / findings from Focus Group 1 (27th May 2021)

Summary of findings: Challenges and difficulties in accessing big data from private digital platforms

- Data transfer can be problematic when it comes to a large amount of data (terabytes)
- (Legal) API restrictions: regulation in EU countries can be different (e.g., German regulation does not provide access to platform data)
- Data Coverage: Self-referentiality of data (e.g., tourism platforms: new users answer with good reviews to already good reviews)
- Reliability of data: Security of data supply: market and platforms are rapidly changing; supply of data can change – this applies even in the context of existing data-sharing agreements
- Geotagged information: Geotagged information are not always available; data affiliation can be unclear (e.g., cities vs. counties) and lead to problems when merging data
- Other challenges can be linked to missing homogeneity between platforms, doublets, or statistical disclosure

Figure 7-1: Results regarding challenges and difficulties in accessing big data from private digital platforms

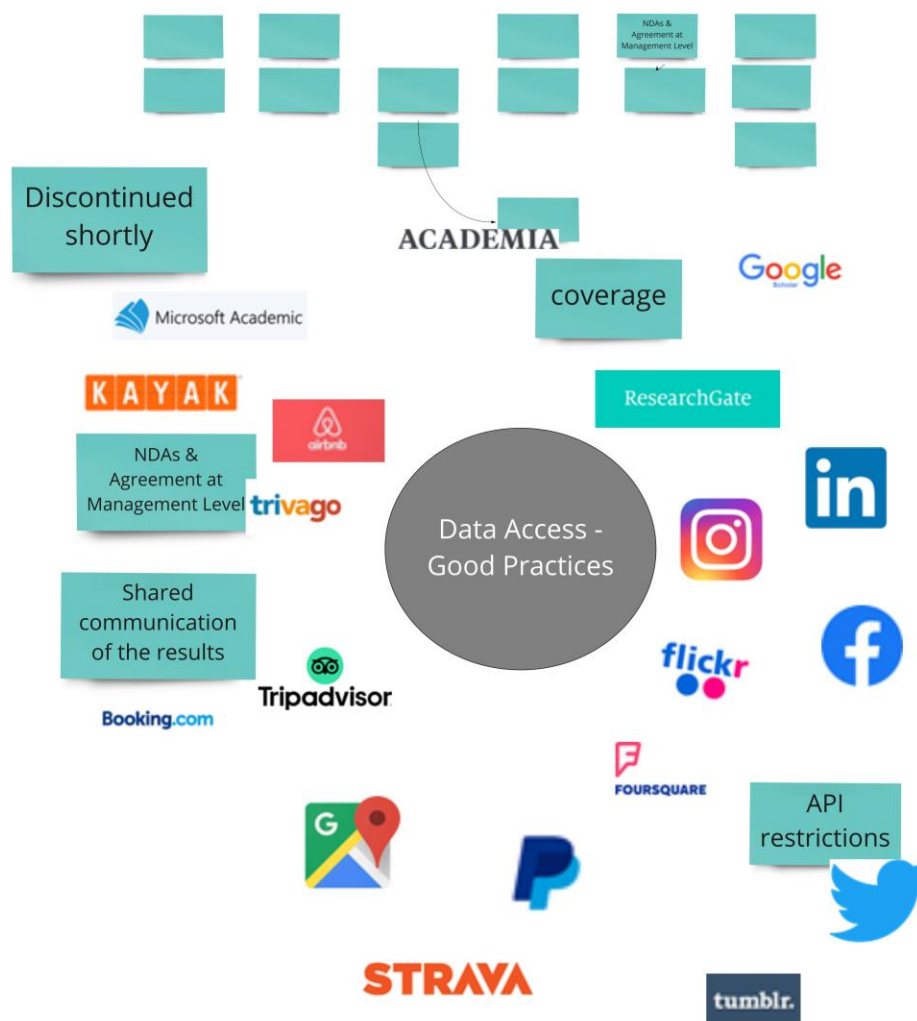


Source: Prognos AG/DevStat (2021).

Summary of findings: Experiences with platforms from the prequalified long list

- Twitter: The UN works with Twitter but only for internal purposes, API access can be, depending on the needed data, a limiting factor. Overall, a lack of used cases has made the work difficult. Overall impression: Difficult to get access
- Nowadays the UN is working a lot with satellite data (AESL data)
- Knowledge platforms: Research data can be used to map the sustainable development goals (SDGs), e.g., very relevant for Pillar 2 “a green Europe” (Territorial Agenda)
- Eurostat started to work on tourism platforms through scraping (in 2014) and then slowly worked towards an agreement

Figure 7-2: Experiences with platforms from the prequalified long list

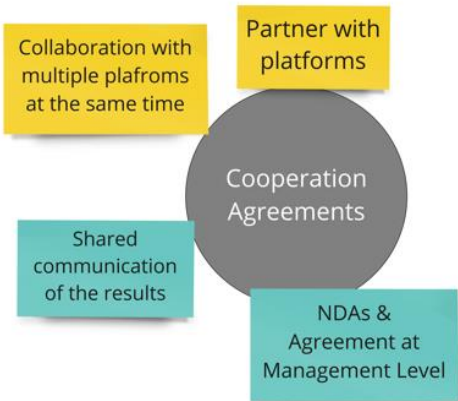


Source: Prognos AG/DevStat (2021).

Summary of findings: Experiences regarding specific cooperation agreements with private digital platforms

- Support from the management level of the platforms is necessary to help with non-disclosure agreements (NDA) and agreements in general
- The market value of the platforms often depends on their data → Partnership with platforms can be difficult because platforms have no real incentive to become a partner

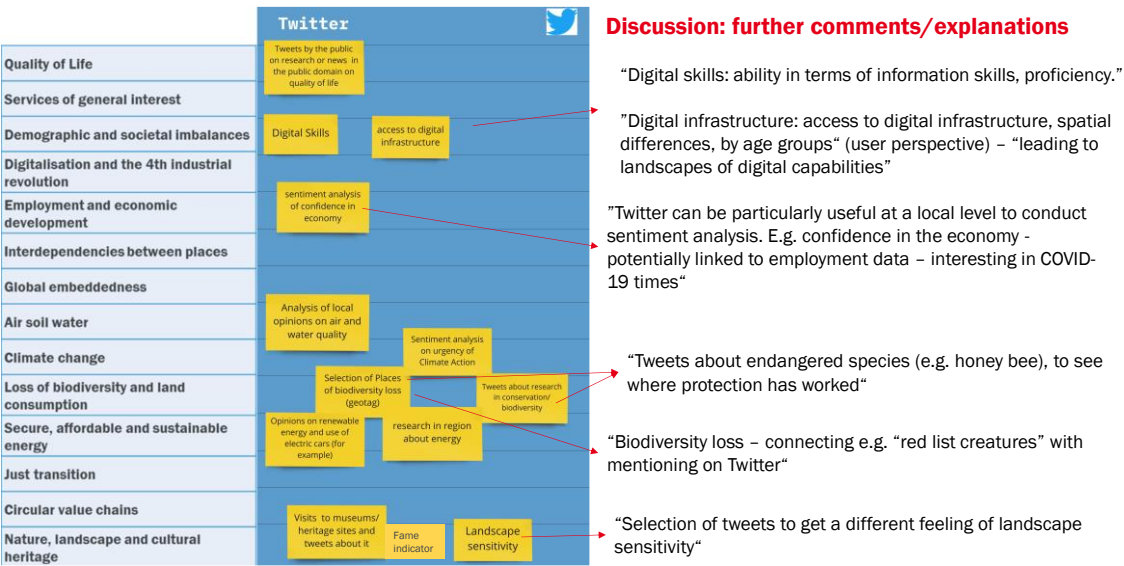
Figure 7-3: Experiences regarding specific cooperation agreements with private digital platforms



Source: Prognos AG/DevStat (2021).

7.3 Results / findings from Focus Group 2 (1st July 2021)

Figure 7-4: What specific indicators could be drawn from big data? - Twitter



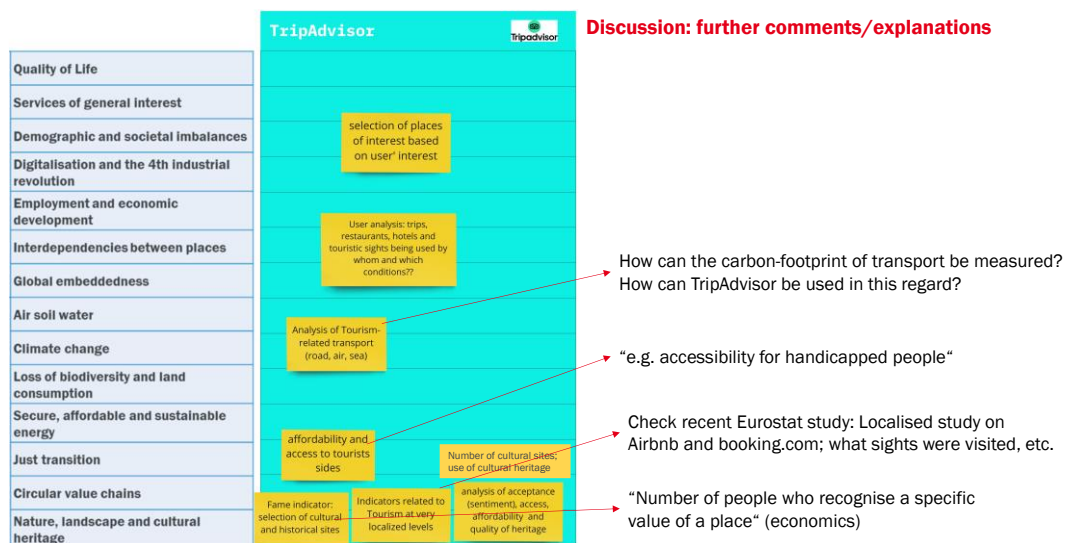
Source: Prognos AG/DevStat (2021).

Figure 7-5: What specific indicators could be drawn from big data? - LinkedIn



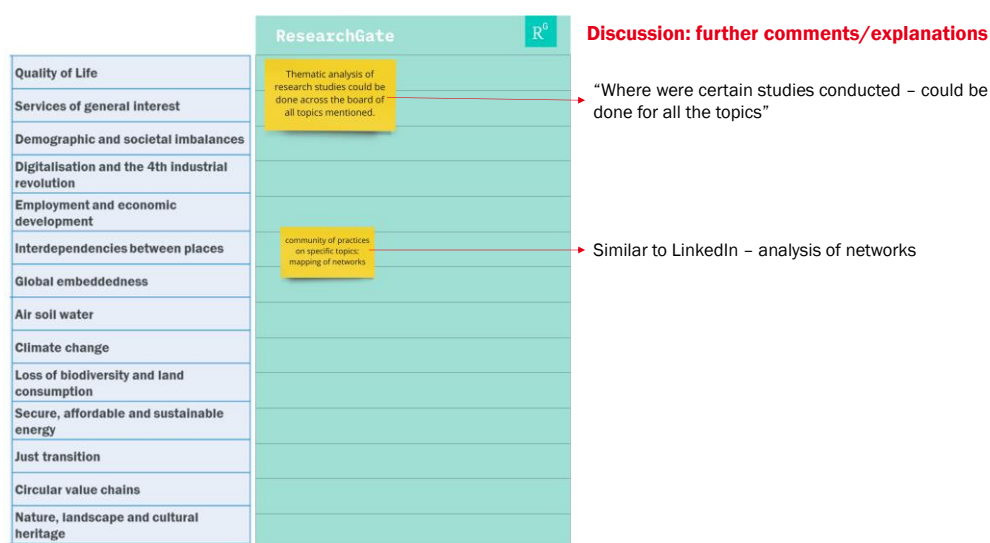
Source: Prognos AG/DevStat (2021).

Figure 7-6: What specific indicators could be drawn from big data? - TripAdvisor



Source: Prognos AG/DevStat (2021).

Figure 7-7: What specific indicators could be drawn from big data? - ResearchGate



Source: Prognos AG/DevStat (2021).

7.4 Experts that participated in the focus groups

Table 7-2: Experts that participated in the focus groups

Albrecht Wirthmann	Eurostat
Andrea Ascheri	Eurostat
Ronald Jansen	United Nations
Lewis Dijkstra	European Commission – DG REGIO
Natalie Rosenski	German Federal Office of Statistics
Natascha Herzog	German Federal Office of Statistics
Oliver Hauke	German Federal Office of Statistics
Jürgen Wastl	Digital Science
Angliioletta Voghera	Politecnico di Torino
Luigi La Riccia	Politecnico di Torino
Eva Schweitzer	German Federal Institute for Research on Building, Urban Affairs and Spatial Development

Source: Prognos AG/DevStat (2021).

References

- Batista e Silva et al. (2017): Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources. <https://doi.org/10.1016/j.tourman.2018.02.020>
- Bowley, R et al.: Insights into career outcomes and skills of Dutch graduates. 2020 Edition.
- Boy, J., Uitermark, J .(2016): How to Study the City on Instagram. <https://doi.org/10.1371/journal.pone.0158161>
- Bertelsmann Stiftung / Jacques Delors Institut (2018): Online platforms and how to regulate them. https://www.bertelsmann-stiftung.de/fileadmin/files/user_upload/EZ_JDI_OnlinePlatforms_Dittrich_2018_ENG.pdf
- Cervera-Ferri, JL (ed.), R. Brennenraedts, D. Fazio, M. Scannapieco, T. Van der Vorst, P. Votta. ESS Big Data Event Rome 2014. Technical Workshop Report. Eurostat https://ec.europa.eu/eurostat/cros/content/2014-big-data-event-final-report_en
- ESPON (2019): Big Data for Territorial Analysis and Housing Dynamics. <https://www.espon.eu/big-data-housing>
- European Commission (2016): Digital Platform for public services. Final Report. <https://joinup.ec.europa.eu/sites/default/files/document/2018-10/330043300REPJRCDigitalPlatformsBM-D2.5FinalReportv051018.pdf>
- European Commission (2019): Measuring labour mobility and migration using big data. <https://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=8264>
- European Commission (2020): Towards a European strategy on business-to-government data sharing for public interest. Final Report prepared by the High-Level Expert Group on Business-to-Government Data Sharing. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=64954
- European Commission (2020): Advanced Technologies for Industry – EU report. Technological trends and policies. <https://ati.ec.europa.eu/reports/eu-reports/eu-report-technological-trends-and-policies>
- European Commission (2020): Advanced Technologies for Industry - Methodological report. <https://ati.ec.europa.eu/reports/eu-reports/advanced-technologies-industry-methodological-report>
- European Parliament Research Centre – EPRS (2016): Big data and data analytics. The potential for innovation and growth. <https://euagenda.eu/publications/at-a-glance-plenary-march-2017-fundamental-rights-implications-of-big-data>
- Eurostat (2021): Inferring job vacancies from online job advertisements. Statistical Working Papers. <https://ec.europa.eu/eurostat/documents/3888793/12287170/KS-TC-20-008-EN-N.pdf/6a86d53e-d0b8-d608-988d-d91f0cef6c21?t=1611673495829>
- Fatehkia, M. et al. (2018): Using Facebook ad data to track the global digital gender gap. <https://www.sciencedirect.com/science/article/pii/S0305750X18300883?via%3Dihub>
- Federal Government of Germany (2021): Datenstrategie der Bundesregierung, p.22. Available online: <https://www.bundesregierung.de/resource/blob/992814/1845634/f073096a398e59573c7526fe-aadd43c4/datenstrategie-der-bundesregierung-download-bpa-data.pdf> (in German; accessed on 08.09.2021)
- Gantz, J., Reinsel, D.: THE DIGITAL UNIVERS IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. In: IDC IVIEW, 2012.
- Griffin, G. and Jiao, J. (2015): Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. <https://ideas.repec.org/p/osf/socarx/5gy3u.html>
- Hartl and Ludin (2021): "Recht der Datenzugänge", MMR, 534 (537).
- Hochman, N.; Schwartz, R. (2020): Visualizing Instagram: Tracing Cultural Visual Rhythms. <https://ojs.aaai.org/index.php/ICWSM/article/download/14361/14210>

- Höpgen, W.; Müller, M.; Fuchs, M.; Lexhagen, M (2020): Flickr data for analysing tourists' spatial behaviour and movement patterns: A comparison of clustering techniques. <https://www.emerald.com/insight/content/doi/10.1108/JHTT-08-2017-0059/full/html>
- Jeng et al. (2016): Information exchange on an academic social networking site: A multidiscipline comparison on researchgate Q&A. doi/epdf/10.1002/asi.23692
- Krugman, Paul R (1991). Geography and trade. MIT press.
- Lenormand et al. (2014): Tweets on the road. Instituto de Fisica Interdisciplinar y Sistemas Complejos
- Loonis, V., & de Bellefon, M. P. (2018). Handbook of Spatial Analysis: Theory and Application with R. Eurostat, INSEE, no. October, 394. Available at <https://www.insee.fr/en/information/3635545>
- Marciano, A. et al (2020): Big data and big techs: understanding the value of information in platform capitalism. <https://link.springer.com/article/10.1007/s10657-020-09675-1>
- Martí et al. (2019): Social Media data: Challenges, opportunities and limitations in urban studies In Computers, Environment and Urban Systems, Vol 74, p. 161-174
- Nemeslaki, András; Pocsarovszky, Károly (2011): Web crawler research methodology, 22nd European Regional Conference of the International Telecommunications Society (ITS2011), Budapest, 18 - 21 September, 2011: Innovative ICT Applications - Emerging Regulatory, Economic and Policy Issues
- Newman, M. E. J., 2001. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E 64, 016132.
- Opsahl, T., Agneessens, F., Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks 32, 245-251
- Sattar, N.S.; Arifuzzaman, S. COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA. Appl. Sci. 2021, 11, 6128.
- Selala, M.; Musakwa, W.: The Potential of Strava Data to contribute in non-motorised transport (NMT) planning in Johannesburg, 2016
- Sen, R. and Quercia, D. (2018): World wide spatial capital. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190346>
- Severo, M. et al. (2015): Twitter data for urban policy making: an analysis on four European cities. https://www.researchgate.net/publication/279175120_Twitter_data_for_urban_policy_making_an_analysis_on_four_European_cities
- Sloan, L., & Morgan, J. (2015): Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. PloS one, 10(11), e0142209. <https://doi.org/10.1371/journal.pone.0142209>
- Spyratos, S. (2017): Using Foursquare place data for estimating building block use. <https://journals.sagepub.com/doi/pdf/10.1177/0265813516637607>
- Strowel, Alain; Vergote, Wouter (2016): Digital Platforms: To Regulate or Not To Regulate? https://ec.europa.eu/information_society/newsroom/image/document/2016-7/uclouvain_et_universit_saint_louis_14044.pdf
- Tostes, A. et al. (n.d.): Studying Traffic Conditions by Analyzing Foursquare and Instagram Data, <https://homepages.dcc.ufmg.br/~thiago/papers/pew10-tostes.pdf>
- Van Aalst, J. (2016): Using Google Scholar to Estimate the Impact of Journal Articles in Education. <https://doi.org/10.3102/0013189X10371120>
- Wang, F.; Xu, Y.: Estimating O-D travel time matrix by Google Maps API: implementation, advantages, and implications, Annals of GIS, 17:4, 199-209, 2011, DOI:10.1080/19475683.2011.625977
- World Bank Group/LinkedIn (n.d.): Data Insights: Jobs, Skills and Migration Trends Methodology & Validation Results. <http://documents1.worldbank.org/curated/en/827991542143093021/pdf/World-Bank-Group-LinkedIn-Data-Insights-Jobs-Skills-and-Migration-Trends-Methodology-and-Validation-Results.pdf>

Xiang et al. (2017): Assessing Reliability of Social Media Data: Lessons from Mining TripAdvisor Hotel Reviews. In: Information and Communication Technologies in Tourism 2017 (pp.625-638)

Zhang, Y.: Data Service API Design for Data Analytics, 2012

Zhu, T. J.; Fritzler, A.; Orlowski, J.: Data Insights: Jobs, Skills and Migration Trends Methodology and Validation Results (English). Washington, D.C.: World Bank Group, 2018.

ESPON 2020

ESPON EGTC
11 Avenue J.F. Kennedy, L-1855 Luxembourg
Grand Duchy of Luxembourg
Phone: +352 20 600 280
Email: info@espon.eu
www.espon.eu

The ESPON EGTC is the Single Beneficiary of the ESPON 2020 Cooperation Programme. The Single Operation within the programme is implemented by the ESPON EGTC and co-financed by the European Regional Development Fund, the EU Member States, the United Kingdom and the Partner States, Iceland, Liechtenstein, Norway and Switzerland.

Disclaimer

This delivery does not necessarily reflect the opinion of the members of the ESPON 2020 Monitoring Committee.