Rheinische Friedrich-Wilhelms-Universität Bonn

Department of Geography / Geographisches Institut

Good Governance Goes Smart – Social Media Data Supporting Evidence-Based Municipal Policies

Development of a Browser-Based Dashboard Prototype for Privacy-Aware Analysis of Location-Based Social Media Data for Application to Urban Planning Illustrated by a Case Study in Bonn

Master Thesis / Masterarbeit

In Partial Fulfilment of the Requirements for the Degree of Master of Science (M.Sc.)

Supervisors:	Prof. Dr. Klaus Greve	Department of Geography Rheinische Friedrich-Wilhelms- Universität Bonn
	Dr. Alexander Dunkel	Institute of Cartography TU Dresden
Submitted by:	Dominik Weckmüller	
Place, Date:	Milan, June 22, 2021	
Matriculation No.:	3219851	

Declarations

- EN: I declare that I have written this thesis independently, that I have not used any sources or aids other than those indicated, and that any parts of the thesis which are taken from other works, either in wording or in meaning, have been marked as borrowed. The same applies to drawings, figures, maps, tables and illustrations. The text part of this thesis (including spaces and footnotes) consists of 191 661 characters.

  DE: Ich versichere, dass ich die Arbeit selbstständig verfasst habe, dass ich keine anderen Quellen und Hilfsmittel als die angegebenen benutzt und die Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen sind, in jedem Fall als Entlehnung kenntlich gemacht habe. Das Gleiche gilt auch für beigegebene Zeichnungen, Kartenskizzen und Abbildungen. Der Textteil der Arbeit (inkl. Leerzeichen und Anmerkungen) umfasst 191 661 Zeichen.

  Milan/Mailand, June 22, 2021 — Dominik Weckmüller

  Place/Ort, Date/Datum — Signature/Unterschrift

- This thesis is written in a gender-neutral way.
- The case study draws from a personal meeting with representatives of the city of Bonn on September 16, 2021 for a preceding research seminar (see BURK, HUHN & WECKMÜLLER 2020; GEOGRAPHISCHES INSTITUT UNIVERSITÄT BONN 2020).
- The entire open-source code base for the dashboard (see WECKMÜLLER 2021a) as well as supplementary material (see WECKMÜLLER 2021b) can be downloaded from GitHub.
- The dashboard development will continue in cooperation with Dr. ALEXANDER DUNKEL and MARC LÖCHNER from TU Dresden for a municipal pilot study. Any future announcements will be published online (see WECKMÜLLER 2021b; 2021c).
- Three live dashboard presentations at different conferences are planned:

	Conference title (reference)	**Date**	**Speaker/s**
1)	"DFNS 2021 – Dresdner Flächennutzungssymposium" (LEIBNIZ-INSTITUT FÜR ÖKOLOGISCHE RAUMENTWICKLUNG 2021)	June 28, 2021	ALEXANDER DUNKEL DOMINIK WECKMÜLLER
2)	"VGIscience 2021 lecture series" (VGISCIENCE 2021)	July 22, 2021	MARC LÖCHNER
3)	"#GeoWoche2021" (DEUTSCHER VERBAND FÜR ANGEWANDTE GEOGRAPHIE E.V. 2021)	October 8, 2021	DOMINIK WECKMÜLLER

Contents

1 Introduction

1.1 Motivation

With more than one billion monthly active users (FACEBOOK 2021a: n.p., as of June 2018), Instagram users are producing an incredible amount of data. Other social media (SM) platforms such as Facebook, Twitter and TikTok impress with similar user statistics. According to ILIEVA & MCPHEARSON (2018: 553) "the vast scale and near-real-time observation are unique advantages of SM [data]" and hence harbor an enormous potential to different application purposes such as urban planning.

Data from and about users are becoming increasingly detailed and, with regard to Instagram as the largest location-based social media network (LBSN), by including spatial information at a rapid pace. In combination with a timestamp and the actual content, this information can be used to make extremely detailed statements about users which is the reason why LBSN harbor an enormous information potential for data science. Within the last 20 years, a broad spectrum of sophisticated SM-specific methods like, e.g. image recognition and natural language processing (NLP) has developed through a combination of data abundance, modern mathematics, more powerful computers and an active international research community. The potential societal and spatial insight that can be gained through LBSM data analysis is considerable which is the reason why LBSM data become increasingly valuable.

At this point however, there is a gap between powerful methods on the scientific side and the citizen side, having no means to access even basic information derived from LBSN.

While the data are relatively easy to query, few research attempts are carried out to make it available to a lay audience despite scientifically well-known societal problems such as spatial injustice, inequity and inequality. Accessing this unprecedented information base, the framework of smart governance (SG) offers the chance to rethink these problems for municipal decision-making.

The aim of this thesis is to build a bridge between the enormous data potential and the people by developing a LBSN-Dashboard and examining a concrete use case on the municipal level of the German city of Bonn for democratic urban planning.

1.2 Problem Statement

Such an undertaking poses certain ethical and legal problems, since the user data belong to the users alone including the right to self-determination over their data on the one hand and the right to privacy on the other. The much too short-sighted but often used argument that the posts were deliberately published is simply not sufficient for an in-depth discussion of privacy and violates the most crucial aspects of privacy.

Even though privacy is in itself a broad topic receiving increasing attention not only in research, but also in society at large (HAZARI & BROWN 2013: 46), the majority of users are not or only partially

aware of what can actually be concluded from what they share or reveal about themselves (KESSLER & MCKENZIE 2018: 6f).

Notwithstanding the definitional vagueness of the concept, data privacy essentially deals with the right to decide autonomously whom to share personal data with and who might know about one's actions (MOORE 2008: 412). Still, privacy is rarely addressed in location-based SM (LBSM) research and, worse, often negligently ignored. In this regard, many negative examples can be found that have analyzed data and published high-resolution results that clearly damage the privacy of users (KOUNADI & LEITNER 2014: 140).

One of the biggest reported scandals in the media in recent years was probably the targeted (ab-)use of Facebook user data by the company Cambridge Analytica. The company exploited social media data to carry out so-called political micro-targeting, i.e. the targeting of individual, unknowing SM users for political purposes in the context of the 2016 United States presidential elections (ISAAK & HANNA 2018: 57).

The interest in valuable information for important societal questions derived from LBSM for smart governance on the one hand and the definitive right to privacy of the user on the other hand must be reconciled.

1.3 Research Interest

For this purpose, this thesis develops a privacy-aware LBSN dashboard based on the HyperLogLog (HLL) algorithm for easy information retrieval and extraction that makes use of the LBSN data richness without recklessly putting user privacy at risk but instead providing a customizable privacy approach.

For this purpose, a dashboard prototype (LBSN-Dashboard) is developed, which is tailored for use in municipalities but offers a high scalability for other purposes or other spatial levels. It focuses on privacy and enables targeted queries that have a high value for smart, municipal decision-making but also directly for the citizens.

A theoretical foundation for embedding the dashboard is laid by attempting to define Municipal Smart Good Governance (MSGG) from the respective subconcepts and then examining the concrete exemplary case for the German city of Bonn in a case study in order to demonstrate the LBSN-Dashboard developed here and embed it in the methodological and theoretical context of this work.

There is a significant need for such a worldwide unique privacy-aware LBSN-Dashboard given the growing socio-spatial inequality, the rapid growth of SM and a growing interest in SM knowledge among municipalities (AGOSTINO 2013: 232). Nevertheless, this interdisciplinary research field of geoinformatics, mathematics, data science, ethics, and law belongs to a science niche.

Due to the limited scope of this work, the aim is not to develop a complete, fully fledged web app, but rather to develop an operational prototype that can be adapted to the individual needs of municipalities and citizens as well as to the respective problem and that can be further developed by an open-source community in the future. However, the dashboard prototype is already ready-to-use within its limited functionality.

On the practical side, the dashboard builds on the previous work of a research team around DUNKEL and LÖCHNER at the Institute for Cartography of the TU Dresden. This includes

- the LBSN framework by DUNKEL et al. (2021),
- the Docker container for the LBSN HLL database (DUNKEL & LÖCHNER 2021a) and
- the corresponding Python package lbsntransform for streaming data into the database (DUNKEL & LÖCHNER 2021b).

The LBSN HLL Database (HLL-DB) uses the HLL implementation by CITUS & CONTRIBUTORS (2021). In the course of data mining, various scripts were created during the preparatory phase of this thesis that efficiently retrieve location IDs from Instagram and the associated posts (see WECKMÜLLER 2021d).

1.4 Structure

This thesis is divided into six chapters and four main parts (fig. 1).

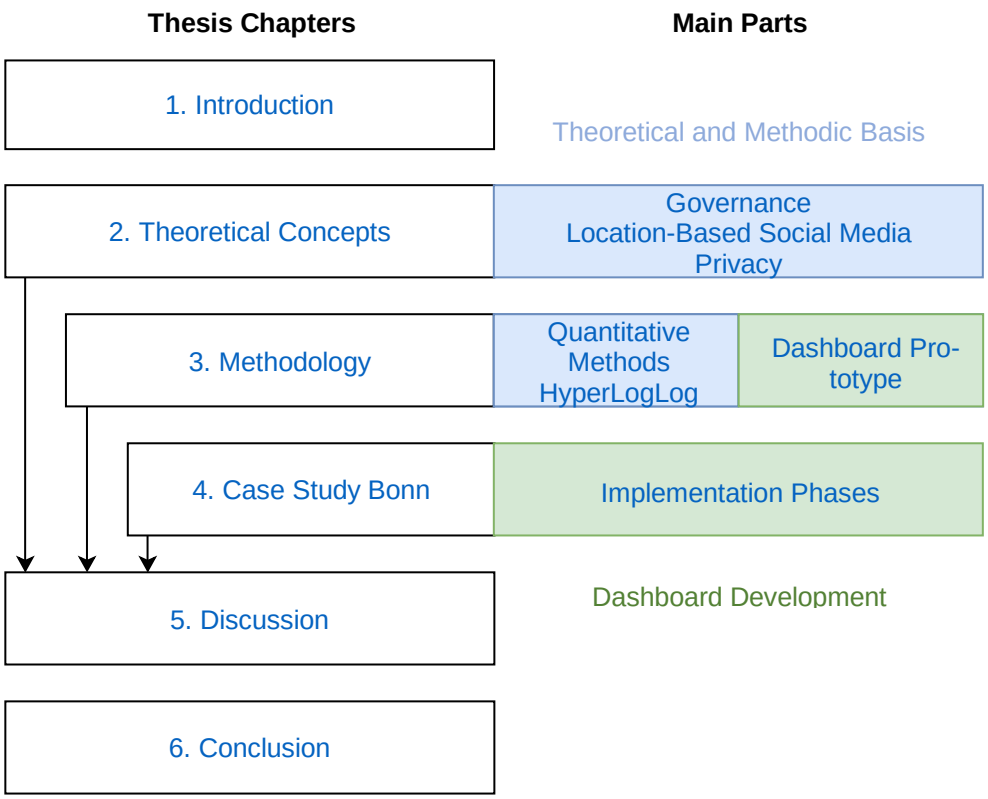


Figure 1: Thesis structure.

Following the introduction (ch. 1), the first main part of the thesis deals with the theoretical foundation of governance (ch. 2.1), which is to be conceptualized with regard to their application at the municipal level to increase well-being and the quality of life of citizens. After the theoretical embedding follows the current state of research of the concepts of LBSM (ch. 2.2), LBSN dashboards (ch. 2.3) and data privacy (ch. 2.4).

In the second part of this thesis, the methods section explains the basic methodological idea behind the dashboard (ch. 3.1). The HLL algorithm is introduced and presented in an application-oriented manner (ch. 3.2). The operators on which the queries in the dashboard are based are explained as well as necessary measures to increase the privacy of the user.

In the third part, following the methodological and pragmatic introduction, the actual dashboard is developed, presented and discussed (ch. 3.3).

The fourth part of this thesis deals with the concrete case study in Bonn (ch. 4), which, after discussion of the data, is divided into three phases and deals with the insights that can be gained through the dashboard.

The findings are summarized (ch. 4.5) with a subsequent outlook about the functions beyond the scope of this thesis which will be aimed at in the future for the pilot study in a municipality (ch. 4.6).

The final discussion (ch. 5) critically evaluates the most important findings with respect to the topics of privacy (ch. 5.1), HLL in practice (ch. 5.2), dashboard implementation (ch. 5.3), and MSGG (ch. 5.4).

The conclusion (ch. 6) sums up the practical findings (ch. 6.1), the case study results (ch. 6.2), the theoretical findings (ch. 6.3) and a suggestion for future research (ch. 6.4).

Since this thesis is very application-oriented and committed to the principles of open source, two GitHub repositories were created for the dashboard development (WECKMÜLLER 2021a) and supplementary material (WECKMÜLLER 2021b).

2 Theoretical Concepts: Governance and Location-Based Social Media

2.1 Governance

Governance is a widely used concept that gained widespread acceptance in research in the 1980s (RHODES 2007: 1246; KJÆR 2004: 1). In the broadest sense, it is understood as “the activity of coordinating communications in order to achieve collective goals through collaboration” (WILLKE 2007: 10), occurring in any group starting from two people only (ibid.)[1].

[1] For a brief introduction see BEVIR (2012), for a very detailed one see KJÆR (2004).

In a political sense, it is understood more concretely as the "process of governing" or the "method by which society is governed" (RHODES 2007: 1246) or any "other constellation of actors" (PEREIRA et al. 2018: 143). Importantly, governance is not to be equated with government, because "governance refers to something broader than government, and it is about steering and the rules of the game" (KJÆR 2004: 7) "in order to enhance the legitimacy of the public realm" (ibid.: 15).

According to HOLTKAMP (2007: 366), governance stands for an analytical, a descriptive and a normative perspective which are often mixed up by the variety of different governance terms. On the one hand, thematically related terms play a major role, as conceptual understandings change. Governance is applied in various areas of life, such as corporate governance (e.g. CHEFFINS 2013), health governance (e.g. YOUDE 2012), social governance (e.g. DEACON et al. 2003) or environmental governance (e.g. YI et al. 2012).

On the other hand, normative additions such as "good" (e.g. SMITH 2007), "bad" (e.g. ROSE & PFEIFFER 2018), "fair" (e.g. BUCKLEY 2009) or "smart" (e.g. PEREIRA et al. 2018) exist, which imply a certain normative, i.e. optimal state that the status quo is to be evaluated against.

Furthermore, descriptive additions denote the geographical scale, e.g., "local" (e.g. HOLTKAMP 2007), "municipal" (e.g. SMEDBY & QUITZAU 2016), "regional" (e.g. FÜRST 2004), "national" (e.g. FISHER 2004) or "global" (e.g. WILLKE 2006). These enumerations are in no way exhaustive and are intended only to express the breadth of governance from which some of this work draws.

It is important to be precise about the terminology and the subdiscourses. The more governance is limited and defined with the additions just mentioned, the more likely it is to have an impact (RHODES 2007: 1246). RHODES (ibid.) even goes so far as to claim that some of these different concepts of governance have "little or nothing in common" and that for effective work with governance, it is necessary to clearly define what exactly the author means by governance.

As "the definitions are used in different subfields of political science, and therefore [...] refer to different debates" (KJÆR 2004: 4), this thesis should not presume to be able to establish a general, universal definition for governance especially across different contexts.

Instead, this thesis focuses on the normative concept of MSGG, which is be explained below, after discussing its subconcepts of GG, SG and MG. Since the goal of this thesis is to make the GG concept fruitful on a municipal level through social media, i.e. to transform it into smart GG, it is also necessary to clarify what "smart" means (ch. 2.1.2) and how it can be conceptualized on a municipal level.

The aim of this chapter is therefore to discuss the different relevant forms of governance and to conclude in a working definition for MSGG.

2.1.1 Good Governance

In order to make the concept of good governance fruitful for this work, a brief historical classification is required in order to understand the core meaning and to grasp the base discourses on which all later discussions build.

GG is a multi-layered concept that was widely introduced into research in the context of development cooperation in the 1990s building on the previous governance debate of the 1980s and driven by the WORLD BANK (see 1991; 1992; NANDA 2006: 269). The WORLD BANK (1991: 1) defines GG as "the manner in which power is exercised in the management of a country's economic and social resources for its development". Thus, it is primarily about a certain form of resource allocation.

Out of "concern for the effectiveness of the development it supports" (ibid.) the WORLD BANK designed this normative concept in order to determine the suitability of recipient countries for receiving development aid on the one hand and to be able to retrospectively legitimize failed projects or a lack of efficiency in the use of financial resources on the other (SANTISO 2001: 3).

GG was initially mainly postulated as deficient, i.e. the concept was used to establish its absence (see WORLD BANK 1991: 6). In the early texts of the WORLD BANK, a lack is mentioned in a variety of ways, e.g. the lack of "democracy" (ibid.: 2), "an adequate legal framework" (ibid.: 6), "educated and trained manpower" and "public accountability" (ibid.), "adequate oversight from national authorities" (ibid.: 9), "transparency" (ibid.: 12), "progress" (ibid.: 18) and "government commitment" (ibid.).

Since GG is "unsettled in its meaning" (NANDA 2006: 269), there is no universally valid definitional consensus. However, certain core ideas can be gleaned from the relevant literature that appear in all definitions.

Table 1: Five principles of good governance (GRAHAM et al. 2003: 3 adapted from UNDP 1997: n.p.).

The Five Good Governance Principles	The UNDP Principles and related UNDP text on which they are based
1. Legitimacy and Voice	**Participation** – all men and women should have a voice in decision-making, either directly or through legitimate intermediate institutions that represent their intention. Such broad participation is built on freedom of association and speech, as well as capacities to participate constructively. **Consensus orientation** – good governance mediates differing interests to reach a broad consensus on what is in the best interest of the group and, where possible, on policies and procedures.

2. Direction	**Strategic vision** – leaders and the public have a broad and long-term perspective on good governance and human development, along with a sense of what is needed for such development. There is also an understanding of the historical, cultural and social complexities in which that perspective is grounded.
3. Performance	**Responsiveness** – institutions and processes try to serve all stakeholders. **Effectiveness and efficiency** – processes and institutions produce results that meet needs while making the best use of resources.
4. Accountability	**Accountability** – decision-makers in government, the private sector and civil society organizations are accountable to the public, as well as to institutional stakeholders. This accountability differs depending on the organizations and whether the decision is internal or external. **Transparency** – transparency is built on the free flow of information. Processes, institutions and information are directly accessible to those concerned with them, and enough information is provided to understand and monitor them.
5. Fairness	**Equity** – all men and women have opportunities to improve or maintain their wellbeing. **Rule of Law** – legal frameworks should be fair and enforced impartially, particularly the laws on human rights.

GRAHAM et al. (2003: 3) summarized these core ideas in their "Five Principles of Good Governance" based on a draft definition from The UNITED NATIONS DEVELOPMENT PROGRAM (1997; UNDP; tab 1.). These principles already suggest that GG is by no means a concept that is limited exclusively to the development context, but has great potential for other contexts. Generally speaking, GG it is about an effective (3. Performance), long-term (2. Direction) poverty reduction and an increase in prosperity for the entire, equal population (1. Legitimacy and Voice & 5. Fairness) based on human rights (5. Fairness) and democratic principles such as participation (1. Legitimacy and Voice), transparency (4. Accountability) and equity (5. Fairness). This UNDP working definition has been built upon by development organizations over time, so that the same core ideas can be found in many definitions.

GISSELQUIST (2012: 23ff) empirically examined 25 working definitions of development institutions from various OECD countries. From these definitions, "seven core components" (ibid.: 2) were identified, which were emphasized across all working definitions. The seven core components follow and extend the above-mentioned definition of the UNDP definition of 1997 (tab. 1):

1) "democracy and representation,
2) human rights,
3) the rule of law,

4) effective and efficient public management,
5) transparency and accountability,
6) developmentalist objectives, and
7) a varying range of particular political and economic policies, programmes, and institutions (e.g. elections, a legislature, a free press, secure property rights)" (ibid.: 8).

Even though these components are derived from development discourses, they serve a great deal to all further discussions. Particularly, some deviations of GG tend to forget about these important core principles as are be described later on.

According to TAYLOR (2016: 2), three main research foci can be identified in the literature: "process-, output-, and outcome-oriented perspectives". The process-oriented "decision-making and implementation processes" focus on the process as such (ibid.: 4), neglecting preconditions and outcomes, whereas the output- and outcome-oriented perspectives put the process in the background and concentrate either on output and outcome. Outputs are concrete things like "laws, regulations, plans" (ibid.: 4), whereas outcome is an abstract category used to judge how sustainable, good, or socially just a policy is. The core idea of GG is to combine "good" things that are self-evident from a Western point of view, but which do not necessarily harmonize with each other (ibid.: 21).

The question with all these definitions, however, is whether GG is a framework for evaluation or a target vision for a state. This dualism has been criticized in literature (see GISSELQUIST 2012: 1; NANDA 2006: 281), generally noting that the ambiguity of the term extends to not just two, but a multitude of possible meanings depending on who is using GG for what purposes. GISSELQUIST (2012: 21) even claims that "[...] good governance means different things not only to different organizations, but also to different actors within these organizations".

However, further points of criticism, which relate in particular to theoretical debates and the problem of the conditionality of receiving development aid are negligible for this thesis.

One particular point of criticism is of utmost importance for this thesis. As GISSELQUIST (ibid.: 21f) expresses the suspicion that a more focused examination of the underlying subconcepts of GG would be useful instead of getting tangled up with GG in a too superficial, general discourse that neglects the previous state of research in the subdisciplines, the subconcepts are concretized later.

Notwithstanding this accusation of old wine in new wineskins, NANDA (2006: 281) sums up that GG is nevertheless useful for analytical purposes. GG is hence used as a general normative framework to embed the following discourses.

To put it in concrete terms, a separate working definition is established throughout this chapter, which, following GISSELQUIST's (2012: 21f) criticism of neglecting the state of research, addresses the respective named subconcepts in the next chapters.

2.1.2 Smart Governance

SCHOLL & ALAWADHI (2016: 22) define SG as "[…] the capacity of employing intelligent and adaptive acts and activities of looking after and making decisions about something" where "the degree of smartness in governance […] is then related to the facilitating capabilities and enabling uses of advanced ICTs" (ibid.: 23). This preliminary definition already implies that the normative term "good" as in GG is not automatically included in the SG concept. WILLKE (2007: 165) acknowledges, that by „redesigning formal democratic governance" the „many achievements of a process of civilization leading to formal democracy and a liberal market economy" must not be given up.

SG has rarely been empirically applied in the literature (e.g. SCHOLL & ALAWADHI 2016) and even more rarely defined in an all-encompassing way.

As with GG, SG is often used interchangeably with other smart concepts such as smart city, smart government, electronic government, electronic governance, or open government, not only in practice but also in theoretical discourse (ibid.: 22). This trend of definitional blurring (ibid.: 24) is exacerbated by, among other things, SG being increasingly used synonymously with ICT governance. SCHOLL & ALAWADHI (ibid.) consider smart ICT governance as a form of SG.

SCHOLL & SCHOLL (2014: 166) state that SG and the concept of smart government are "closely related," with smart government "resting on the foundation" of SG. The lack of empirical studies as well as the conceptual fuzziness complicate the work with the concepts (ibid.). Some of the criticisms of the smart cities debate can, or rather must, be applied to SG. The simplest accusation of a hype around the title smart, especially for marketing purposes (HOLLANDS 2008: 305), does not seem far-fetched in today's global neoliberal competition between cities (see ANTTIROIKO 2015) and can easily be confirmed with a look at the websites and initiatives of cities. One can hardly find a city that does not try to adorn itself with the fancy labels smart or digital. A look at the smart city website of the city of Bonn, for example, reveals all the important keywords: "smart", "digital", "open" and "participation" (BUNDESSTADT BONN 2021a: n.p.).

Equally problematic is the fact that in general an extremely uncritical attitude towards the concepts is taken. Unfortunately, this thesis does not offer the necessary scope to discuss whether i.e. the liberal market economy as mentioned by WILLKE (see 2007: 165) is best for a smart society or what is exactly meant by not giving up the "achievements […] of civilization" (ibid.).

HOLLANDS (2008: 305) asks a trenchant question: "Which city, by definition, does not want to be smart, creative and cultural?". What this question alludes to is the fact that different actors can associate different measures with smart city, smart government or SG or, like WILLKE (2007: 165), have a concrete normative idea of SG that is not necessarily shared equally by all societal actors. In this context, priorities may be set differently and, in extreme cases, individual attributes may differ from one another. For

example, an increase in efficiency in administration can probably easily be achieved at the expense of participation. However, the reduction of participation fundamentally contradicts the idea of GG.

In terms of economic (neo-) liberalism, equity and equality expose false theoretical promises. A frequently cited quotation by ROBINSON (2016: n.p., para. 16; e.g. as cited in BARNS 2018: 2) draws an apt, dystopian example scenario in relation to smart cities, which can also be applied to SG:

> "[…] the massive investments that are being made in smart technology at a scale that is transforming our world are primarily commercial: they are investing in technology to develop new products and services that consumers want to buy. That's guaranteed to create convenience for consumers and profit for companies; but it's far from guaranteed to create resilient, socially mobile, vibrant and healthy cities. It's just as likely to reduce our life expectancy and social engagement by making it easier to order high-fat, high-sugar takeaway food on our smartphones to be delivered to our couches by drones whilst we immerse ourselves in multiplayer virtual reality games."

The contrast between commercial intentions and citizens' interests becomes extremely clear and is unfortunately already found in similar forms in reality (see EUBANKS 2018). It is therefore all the more important to establish a valid working definition for SG that has the necessary sharpness and does not pander to certain individual interests, but rather understands itself entirely in terms of an empowered population[2].

The criticisms mentioned here must be tackled for a profound conceptualization of a normatively "good" SG but are not intended to question the overall utility of the concept. In fact, PEREIRA et al. (2018: 143) sees a societal disruption justified by the digital transformation to which SG can provide answers. WILLKE (2007: 165; cf. 2006) concretizes that, in addition to globalization, this change from an industrial to a knowledge society also poses a challenge to SG.

PEREIRA et al. (2018) do an excellent job of summarizing the current state of scholarship around SG and identify an emerging form of SG that this thesis uses. They untangle the definitional chaos and establish a transformational framework that has smart city governance as its goal (ibid.: 153f).

[2] The debate on technological citizenship deals with participation in a modern society shaped by technologies (cf. FRANKENFELD 1992, cf. VALKENBURG 2012) and could be linked to this discussion.

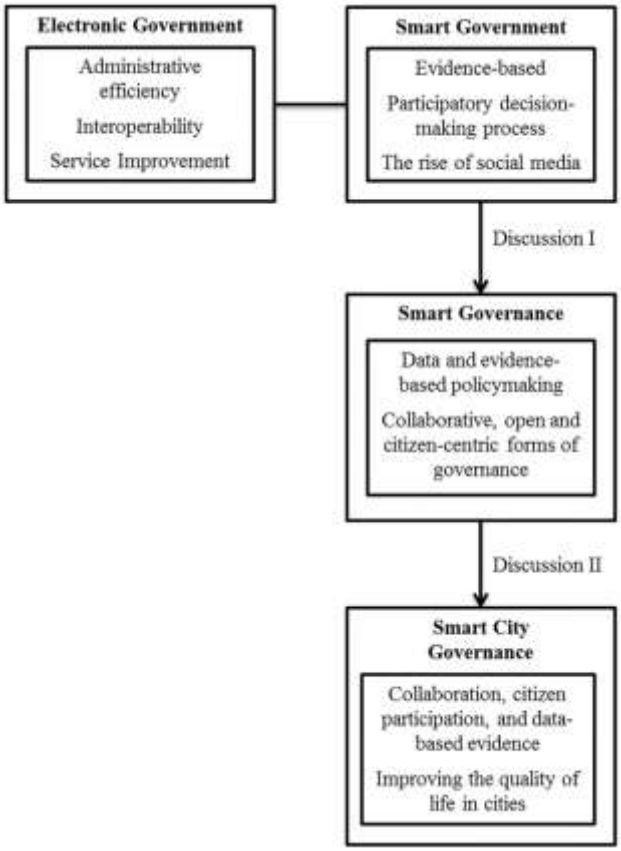


Figure 2: Building smart city governance (PEREIRA et al. 2018: 153).

They state that technology, or ICTs, are the connecting piece of smart government and SG (ibid.: 155; fig. 2). New digital platforms, and SM in particular, generate pressure for change and carry the potential to promote participation, empowerment, transparency, openness (ibid.: 150). In order to promote evidence-based policymaking (ibid.: 149), data-driven decision making is necessary. SM play a special role in this context, as they are increasingly used by governments to introduce forms of "citizen-centric governance" (ibid.: 150). To what extent a SM dashboard can contribute to making a municipality smarter is examined in the case study (ch. 4).

According to PEREIRA et al. (ibid.: 153; fig. 2), the following criteria are the key to smartness:

- data and evidence-based decision-making
- participation
- collaboration
- openness and transparency
- citizen centricity
- ICT-promoted (i.e. through SM) transformation

These build on electronic government, i.e. administrative efficiency, interoperability and service improvement (ibid.).

WILLKE (2007) makes use of the basic idea of this chapter to symbiotically combine several different forms of governance in order to apply them in a targeted manner. He links the concept of SG described here with GG on a global level (ibid.: 7f).

In the following chapter, the criteria developed here according to PEREIRA (see 2018: 153f) are processed into a similarly specific from of governance, following WILLKE (see 2007: 7f). Instead of global governance, MG is introduced here as the last form of governance and outlined with regard to GG and SG.

2.1.3 Municipal Governance

MG is considered a subform of local governance (see TAYLOR 2016) or urban governance (see FOSTER 2006). Local governance deals with the conceptualization of GG on a certain local level.

MG refers to the level of the municipality and its citizens (see SMEDBY & QUITZAU 2016) and represents the level at which GG can be made most operable and feasible (DAMKOWSKI & RÖSENER 2004: n.p., ch. 2.3). A wide variety of definitional attempts also circulate in MG, which, much like its sister concepts GG and SG, pose a "thorny problem" (DOLLERY & JOHNSON 2005: 2).

MG is concerned in particular with planning activities, policy-making and the way in which a municipality interacts with its citizens. In direct comparison to global governance, a high level of personal trust often plays a role here, which helps to reduce corruption and increase "levels of participation and policy compliance" (TAYLOR 2016: 14).

Good local governance has been concretized and made fruitful for practical application by the World Bank and the Bertelsmann Foundation, among others (e.g. SCHÖLER & WALTHER 2003).

HOLTKAMP (2007) notes a number of advantages of good local governance. He is convinced that, on the one hand, good local governance should develop more problem-adequate solutions and new ideas through the use of social knowledge (ibid.: 374). On the other hand, state resources could be supplemented by social actors and implementation resistance and implementation times of large-scale infrastructural projects could be reduced (ibid.). This, in turn, would lead to greater legitimacy and better opportunities for participation, and thus to a reduction in the much-cited disenchantment with politics and increased responsiveness on the part of politicians (ibid.: 375). Furthermore, this increases the transparency of political decision-making processes (ibid.).

The important thing, he argues, is not to narrow down on one governance type, but to achieve a balanced governance mix (ibid.: 375). The governance mix for this thesis, is explained in the following.

2.1.4 Municipal Smart Good Governance

Based on the previous discussion of the concepts GG, SG and MG, a preliminary working definition for MSGG as a normative and descriptive framework is to be established.

Table 2: Key aspects of good, smart and municipal governance.

	Good Governance	Smart Governance	Municipal Governance
Key Aspects	• Participation • Democracy / Consensus / Responsiveness • Human Rights • Accountability/Equality • Effectiveness / Efficiency • Transparency • Equity • Reduction of Poverty / Increase in Well-being	• Participation • Collaboration • Citizen Centricity • Efficiency • ICT-promoted Transformation • Interoperability • Transparency • Openness • Service Improvement	• Participation • Consensus • Collaboration • Efficiency • Transparency • Locality / Target-orientated • Trust / Low-threshold • Quality-of-life-orientated
Reviewed Literature	GISSELQUIST (2012); NANDA (2006); GRAHAM, PLUMPTRE & AMOS (2003); SANTISO (2001); UNDP (1997); WORLD BANK (1991; 1992)	BARNS (2018); PEREIRA et al. (2018); SCHOLL & ALAWADHI (2016); SCHOLL & SCHOLL (2014); HOLLANDS (2008); WILLKE (2007)	SMEDBY & QUITZAU (2016); TAYLOR (2016); HOLTKAMP (2007); FOSTER (2006); DOLLERY & JOHNSON (2005); DAMKOWSKI & RÖSENER (2004); SCHÖLER & WALTHER (2003)

Tab. 2 summarizes the most important elements of the individual definitions. The concepts show large areas of overlap. In particular, there seems to be a universal consensus regarding the basic values of a democracy, such as human rights, the rule of law and participation along different contexts which could be interpreted as the "good" things in GG (TAYLOR 2016: 21) or the "many achievements of a process of civilization" (WILLKE 2007: 165).

GG provides the normative roots against which MSGG must measure itself. SG emphasizes the indispensability of the crucial medium of ICT and MG the municipal, practical context of application. Accordingly, the working definition of MSGG includes the key aspects from above (tab. 2). In short, for this thesis MSGG shall be defined as follows:

> **MSGG is the capacity of improving municipal democratic decision- and policy-making processes and outcomes through the transparent and open usage of ICTs such as LBSM in order to increase citizens' quality of life and well-being.**

2.1.5 The Lack of Space

Although the adjectives “global”, “local” and “municipal” of the governance concepts discussed here address spatial concepts, they are insufficiently addressed by the relevant literature (e.g. DOLLERY & JOHNSON 2005). It almost seems as if the so-called “spatial turn” (see DÖRING & THIELMANN 2015) has passed the relevant professional debates by, when space is only thought in geo-deterministic patterns.

In other publications, “spatial”, e.g. “spatial governance”, is regarded as a nice accessory, catch phrase or title decoration, but they completely disregard what it means to deal with space. Often, spatial governance only means governance or GG (e.g. ALLMENDINGER & HAUGHTON 2013 can be read equally well without “spatial”) or spatial is related to spatial planning, so that spatial planning and governance becomes spatial governance (e.g. ALLMENDINGER 2016).

Be it due to conceptual ambiguity, the literature generally neglects space. Still, it is of utmost importance to be aware of the concepts of space through which one argues and that these can have an enormous impact through their interpretative power (SOJA 2015: 241) when space is deconstructed, for example, through poststructuralist approaches (STRÜVER 2011: 671).

Therefore, in order to highlight the necessity and potential of LBSM in the following and to enrich the debates mentioned above, the four main spatial concepts of geography are briefly addressed.

The main conceptions of space are the absolute notion of physical space, as a simple “container” (GEBHARDT & REUBER 2011: 647), the relational space, e.g. through social contacts, the perceptual space, in which impressions of the space determine it, and the constructed space, which is (un)consciously created by the acting human being (GEBHARDT & REUBER 2011: 647; GÜNZEL 2010: 193; WERLEN 2008: 365f).

Container space has been and is often used in the context of the scale of governance. “Municipal” translates for some authors simply as “on a municipal level”, although they also address relationships, perceptions and constructs in their work. However, the local specifics of networks of relationships and effects, political difficulties, etc. is what constitutes this space in the first place. The communal space with all its structures of action simply does not exist without these components. These aspects are seldomly ignored in literature but if separated from space, certain spatial interrelationships cannot be (spatially) investigated.

If the key aspects of GG, SG and MG (tab. 2) are rethought in connection with space, entirely new possibilities open up. Within the scope of this thesis, not all of them can be dealt with in detail. Two important concepts that can be fertilized by space are explained here.

In the spirit of RHODES (2007: 1246), who argues for a conceptualization of any governance concept, including its subdiscourses, equality and equity as important governance pillars are further discussed.

2.1.6 Spatial Equality and Equity

Equality and equity are fundamental components of a democracy that seek to balance entitlements and resources between different groups (JOFFE-NOTIER 2020: 109f). In essence, it is about people being able to participate in society in a similar way and being entitled to certain resources.

Equality and equity are expressed through resource allocation in a direct (direct payments, e.g. child benefits or unemployment benefits) or indirect (local recreation areas, public transport connections) way. The principles of equality and equity are to be used by the dashboard of equitable resource allocation in order to be able to plan specific outputs such as concrete measures or new infrastructure in relation to the various analysis foci according to TAYLOR (2015: 2ff) on the one hand and to strengthen the outcome, i.e. a smarter, better municipal governance, which is measured against the GG principles listed above on the other hand.

There is a lively debate in literature about equality and equity, not only because they are "widely confused" (BRONFENBRENNER 1973: 9). Put simply, according to BRONFENBRENNER (ibid.) equality is "basically objective" as the amount of a certain good per person can be measured whereas equity is "basically subjective" as it is subject to "ethical judgment".

In Germany, after the fall of the Berlin Wall in 1989, a lively debate on equal living conditions (DE: "gleichwertige Lebensverhältnisse") began in the field of spatial planning and development, which was not only transferred to an equalization between East and West but increasingly to small-scale levels such as states, regions and municipalities (MÄDING 2021: 73f). In the context of this debate, it seems even more important to make evidence-based decisions, e.g. about what exactly is unequal or where exactly certain infrastructure is missing. For such policies a dashboard can contribute significantly by providing a solid information base.

While laws in the sense of equality, for example, apply equally to all, an equal distribution of resources turns out to be more difficult to grasp. Without slipping too much into moral philosophical debates (cf. SEN 1980), which cannot be further discussed, equity deals with the need, i.e. the concrete diverse needs of individuals or population groups (LUCY 1981: 448f).

When defining equity more towards the concrete spatially relevant need of a person, it becomes obvious that certain resources are more in demand by certain groups and less by ohters. This idea is illustrated well by an example, because "[...] equidistant access to kindergartens is not useful when the population's age distribution is considered. Not every district in a city has a similar need for kindergartens" (DUNKEL 2016: 71).

Some authors therefore attribute a special significance to this need (see FOLGER, SHEPPARD & BUTTRAM 1995), which constitutes a separate category of analysis.

The lack of information about different, spatially distributed needs, which is difficult to capture in certain contexts, is a major challenge, especially in the context of uncertainties (DWORKIN 1981: 187) provoked by changing, hard-to-evaluate needs and a scarce information base. These uncertainties are also emphasized by WILLKE (2007: 177) arguing, that there is no universal authority that can ethically evaluate needs and determine that a need, for example, public drug use in certain places should or should not be legalized. This research field is complex and full of tension between ethics, morality, politics and geosocial, democratic negotiation processes, but are not discussed further, as the dashboard is only intended to provide the information basis for all stakeholders.

If spatial equality and spatial equity are turned into the opposite, a phenomenon arises that has meanwhile gained momentum thanks to popular authors. In his 1973 standard work "Social Justice and the City", HARVEY (2010: 101) already postulated the problem of spatial resource allocation in the sense of maximizing social justice. To this end, he conceives the principle of "territorial distributive justice" as a balance between "need, contribution to common good, and merit" (ibid.). In order to adequately address the "variations in demands and needs in the population" however, it is theoretically necessary to have "information concerning the utility scales of each individual in the population" (ibid.: 90). He thus acknowledges the necessity and indispensability of spatial information.

SOJA (2013: 47) builds on this concept and speaks of "spatial justice" or "spatial injustice [...] as the outcome of countless decisions made about emplacement, where things are put in space". SOJA (ibid. 60f) links the production of spatial injustice to the "right to the city" and claims that many institutions such as the EU already exist to specifically combat spatial injustice. However, those institutions in turn require spatial information that is specifically processed. In order to counter the "discriminatory geographies", capacities i.e. "geographical information systems" (ibid.: 51) have already been used in the past. However, such systems can only promote spatial justice in combination with "will and awareness" (ibid.).

As for example "socio-economic and ethnic divides are entrenched and socially reproduced through physical space and everyday life" very much in a Foucauldian sense, SM data are "a useful asset for social-equity research in cities" which GIS researchers already have investigated (ILIEVA & MCPHEARSON 2018: 555).

Spatial information thus seems to play an important role in enabling equity and ensuring evidence-based decisions and policies after targeted democratic negotiation processes. A LBSN dashboard can serve as such a spatial information system and data hub particularly for the spatial component of spatial equality and spatial equity.

This works aims at concretizing the discourse on its underlying key aspects on the practical example of LBSM in Bonn via a privacy-aware dashboard. In order to inform MSGG spatially by LBSM, the necessary missing concepts are discussed below.

2.2 Location-Based Social Media

The terms LBSM[3] and LBSN were coined in the late 2000s (see FISCHER 2008) with the exponential growth of SM (EVANS & SAKER 2017: 64). As TSOU & LEITNER (2013: 55) put it

> "GIScientists can now trace, monitor, and map the spread of social movements, protests, disease outbreaks, nature hazards, elections, political campaigns, etc. in cyberspace by digitally collecting social media and online content."

FISCHER (2008: 5) discusses theoretical implications of a virtual construction of space and recognizes in line with the fourth understanding of space as a construct (ch. 2.1.5), that space would be created through the "self-organized constructive process of communities attaching subjective everyday knowledge to geospatial representations". This perspective assigns LBSM not only a passive but an active, constructive role and also changes the perception of physical space and, as often emphasized, the social interaction in that space. If publicly accessible, despite legal and ethical concerns (ch. 3.3.3), LBSM data can partly be considered volunteered geographic information (VGI, FISCHER 2008: 4), i.e. voluntarily published spatial information.

The spatial component makes the decisive difference to conventional SM. It can exist in different forms. Either as a location with exact coordinates, as a location reference (i.e. city, landmark or point of interest (POI)) or as an undefined geo tag (#bonn). These different spatial references are discussed in more detail in the methods section (ch. 3.3.1.1).

LBSM have been a constant subject to research in most different fields. Prior to the 2018 Cambridge Analytica scandal disclosures, user and post data could be downloaded from the largest SM platforms (Facebook, Instagram, and Twitter) via the publicly available application programming interfaces (APIs) by anyone with certain limits. This data ubiquity fueled science but also triggered a certain privacy carelessness (ch. 2.4).

Thus, a wide range of topics can be identified in the literature using LBSM data. Research using raw data includes health (e.g. PADMANABHAN 2014), tourism (e.g. LIM et al. 2019; HASNAT & HASAN 2018; SHANG et al. 2016; ABBASI et al. 2015), spatial identity (e.g. SCHWARTZ & HALEGOUA 2015), gender studies (e.g. YUAN, WEI & LU 2018), city development and urban sprawl (e.g. JIANG & MIAO 2015), socio-spatial behavior (e.g. GAO, TANG & LIU 2012), urban movement dynamics & urban human mobility (e.g. CAO et al. 2015; HASAN, ZHAN & UKKUSURI 2013) and event reactions (e.g. DUNKEL et al. 2019).

[3] For a detailed introduction to LBSM see EVANS & SAKER (2017).

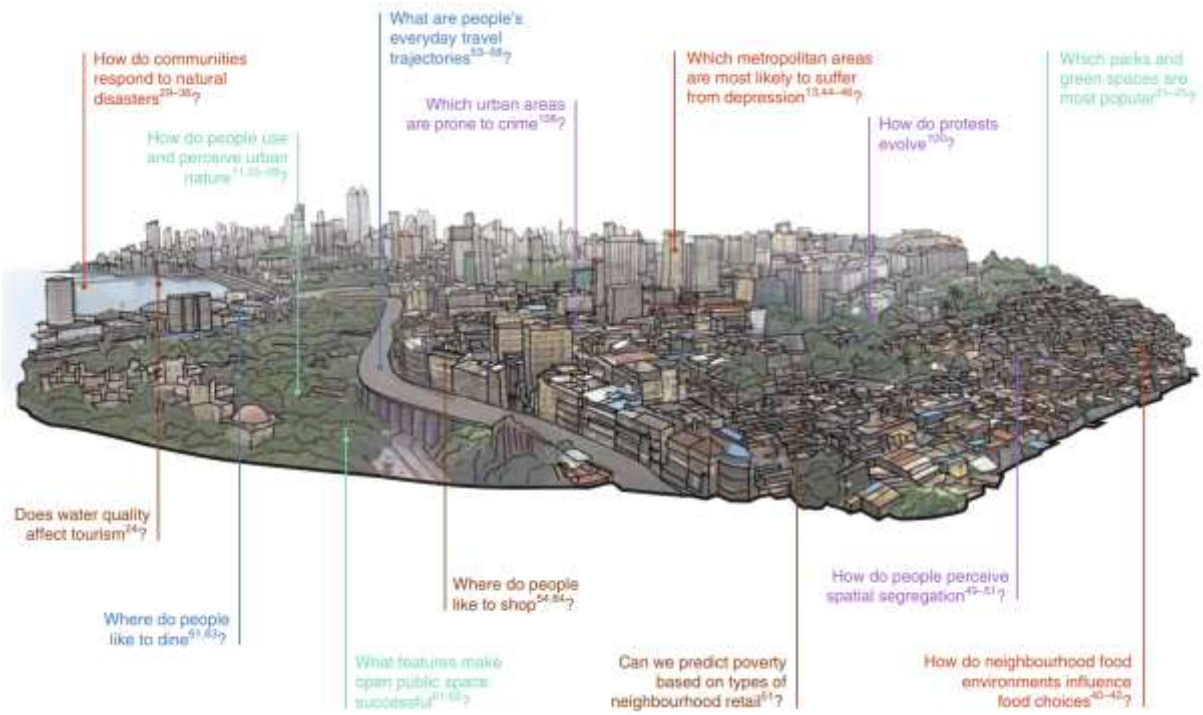


Figure 3: The wide range of emerging opportunities for urban-sustainability research provided by big data from social media (ILIEVA & MCPHEARSON 2018: 554)[4].

ILIEVA & MCPHEARSON (2018:554) sum up this wide range of already ongoing research fields on an urban scale and regard SM data as promising for "addressing key questions" (fig. 3). Such questions and urban struggles can be tackled only with an appropriate information base which the dashboard could supply.

DUNKEL et al. (2019) provide a conceptual LBSM framework in the context of reactions to events based on which SM posts can be classified into different categories for analysis purposes and which is also applies in the case study (ch. 4).

They divide a post into four facets (ibid.: 786f):

- Temporal: Time or period
- Spatial: Location or spatial relation
- Social: Demographic makeup
- Thematic[5]: Meaning and content

[4] For the referenced literature see the original paper.
[5] The thematic facet is also referred to as "topical" (e.g. LÖCHNER, DUNKEL & BURGHARDT 2018: 2). They are used interchangeably.

Figure 4: Abstraction layers for each facet (LÖCHNER, DUNKEL & BURGHARDT 2018: 2 based on DUNKEL et al. 2019: 784).

The facets can be divided into different granularities (fig. 4) and are used in this thesis for three reasons. First, it provides an analysis scheme proven in the literature in similar studies to examine individual posts more closely (ibid.: 784). Second, it allows different LBSMs and their posts to be compared and brought to a common level of analysis, which is an important factor in the context of an SM dashboard (DUNKEL et al. 2021: n.p.). Third, there is an already existing practical implementation offering an easy setup as well as solid database for the application backend (see ibid.).

In recent years, with respect to legitimate concerns of lack of privacy and data protection, most public SM APIs have been withdrawn from public access, with certain exceptions[6]. Especially in combination with spatial information, which enriches the data enormously, it is essential to take privacy seriously (SCHWARTZ & HALEGOUA 2015: 1657). Researchers now find themselves in a certain area of tension between the enriching possibilities listed above in the interplay of LBSM and GIS as a "golden era" (JIANG & MIAO 2015: 303) on the one hand and the concern about legal and ethical concerns on the other.

2.3 The Lack of Location-Based Social Media Dashboards in Planning

However, raw data pose a greater problem in practice, outside of pure research. Even though „[i]nteractive online dashboards are an accessible way to summarize complex information to the public" (PELLERT et al. 2020: 4), very few dashboard-like tools exist that make use of LBSM in the areas of health (e.g. PELLERT et al. 2020; PADMANABHAN 2014), in a disaster context of storms and floods (e.g. TSOU et al. 2015;) or in a geomarketing context for site analysis (e.g. ANDERSON et al. 2019; LIN et al. 2016), all acknowledging the fact, that "[w]rangling APIs, scraping, and analyzing big swathes of data is a skill set generally restricted to those with a computational background […]" (BOYD [sic] & CRAWFORD 2012: 674) and should be made more accessible.

The COVID-19 pandemic health dashboards recently constituted the main focus of public institutions

[6] Facebook's Graph API has been closed in its original form, but can still be queried very easily. For this purpose, a separate Instagram downloader was programmed for this thesis (WECKMÜLLER 2021d), which can quickly query the data. In addition, there are developer accounts to which researchers can apply, e.g. for Twitter. Flickr's API is still publicly available.

(PELLERT et al. 2020: 4) and might deliver a general notion about the utility of data dashboards to the people.

However, at the time writing, there is no publicly available SM dashboard in the world that makes privacy-aware LBSM data from different platforms bundled together accessible to laypersons for different contexts and that takes concerns about privacy, law and ethics seriously.

ILIEVA & MCPHEARSON (2018: 553) point out the increasing need for such an attempt:

> „Global urban science remains fragmented and disconnected from global and local policy and planning, highlighting the need for new tools and data to advance understanding of complex urban dynamics, and to support decision-making for sustainability transformations."

If an application is to work live or near-real time, data must be stored somewhere to be processed and made available again. Here, not only the justified concern about data privacy and ethics plays a major role, but also very practical legal hurdles, since the large SM platforms generally do not allow competing platforms based on their user's data as e.g. INSTAGRAM (2021a: n.p.) states: "**You can't modify, translate, create derivative works of, or reverse engineer our products or their components.** [emphasis in original]" – even though the data actually belong to the users as the terms of use state: "**We do not claim ownership of your content, but you grant us a license to use it.** [emphasis in original] Nothing is changing about your rights in your content." (ibid.).

So, on the one hand, the challenge for such a dashboard and its active use in planning operations is not to work with raw data so that the data cannot get lost and that the public implementation remains legally and ethically unproblematic. The challenge of user privacy, which must not be exposed in the dashboard, is further elaborated below.

2.4 Data Privacy

Data privacy developed at a rapid pace in the 2000s. It used to be a rather vague concept despite the fact that it has become part of most people's everyday lives (BARKER et al. 2009: 42).

Generally, as MOORE (2008: 412) sums up, literature agrees on common aspects of privacy, i.e. a right to decide autonomously whom to share personal data with and who might know about one's actions where the default is always beneficial for the data owner meaning no information at all for third parties. Contrary to a missing clear-cut definition, the data privacy goal within the scope of this thesis is pretty clear, i.e. to "release statistical information about the population who have contributed to the data without breaching their individual privacy" (GEHRKE, LUI & PASS 2011: 432).

In more recent literature, the conflict between user privacy and growing commercial interests is elaborated well. LI, SHARMA & MOHANTY (2020: 10) claim that disclosed data can cause serious harm to individuals but "driven by economic advantages" companies capture, store and use data from many

different sources and turn these to knowledge through analysis. Since knowledge means profit, LBSM companies have a high incentive of storing, aggregating and collecting more data about their users. However, the "frequent incidents of personal data leakage" (ibid.) such as the scandal about Cambridge Analytica prove the controversy of such commercial practices. As BAIK (2020: 2) sums up recent research, a general shift can be observed from privacy as universal right or dignity towards a good or a commodity. The problem here, is that "the creator of data – a user – is not necessarily equal to the monetary beneficiary of the data – a digital platform" (ibid.).

At this point, a common possible misconception is that privacy is automatically given if the individual data are only encrypted well enough and are not accessible to the public. Instead, SWEENEY (2002: 561) claims "computer security is not privacy protection". Privacy does not end with authentication and access control, but begins with the data themselves (ibid.). Put simply, privacy must be granted for the data owner at the stage where a third party has access to the data or in other words: "anything that can be learned about a respondent from the statistical database should be learnable without access to the database" (DALENIUS 1977: n.p., as cited in DE CAPITANI DI VIMERCATI et al. 2012: 13).

The strictest of definitions still allowing a reasonable degree of analysis is called differential privacy (DP). Put simply, "differential privacy ensures that the removal or addition of a single database item does not (substantially) affect the outcome of any analysis" (DWORK 2008: 2) by adding some noise to the data. However, DP methods induces certain limits to usability in practice (DUNKEL, LÖCHNER & BURGHARDT 2020: 3) which is not further discussed here.

The idea of DUNKEL, LÖCHNER & BURGHARDT (2020) follows a different approach, relying on a combination of system-immanent mitigation strategies mainly focused on HLL (see ch. 3.2) and classic computer security approaches including anonymization and encryption.

2.4.1 Privacy Models

There are plenty of privacy models and definitions usually tailored to the needs of specific purposes. For the scope of this thesis, such a custom privacy model particularly suited to a privacy-aware real-life HLL-based LBSM-dashboard needs to be tailored in the following as nothing comparable has been done before.

The privacy model of XU et al. (2014: 1151) focuses on the actors in a data and application flow. They distinguish privacy based on actors in a data processing, i.e. between

- "Data Provider", e.g. individual SM user,
- "Data Collector", LBSN e.g. Instagram,
- "Data Miner", e.g. a research team or me within the scope of this thesis and
- "Decision Maker", the users of the LBSN Dashboard e.g. office of urban planning (ibid.).

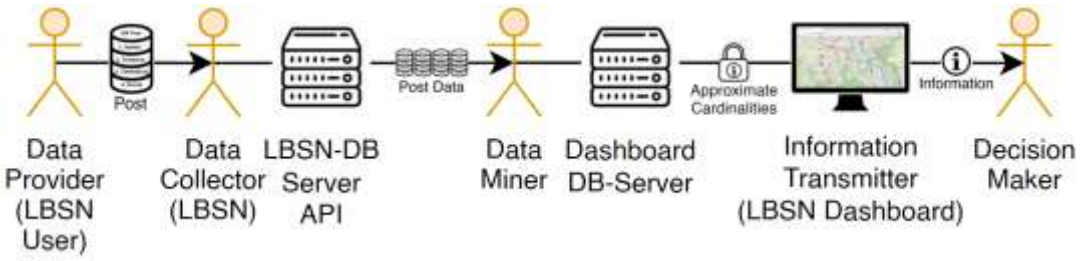


Figure 5: A simple illustration of the application scenario with data mining at the core (modified after XU et al. 2014: 1151).

Fig. 5 describes this processing chain, where the Information Transmitter is the actual LBSN dashboard. For each of these actors, certain privacy conditions apply, each of which is predefined by the particular data stakeholders. For a real-life application one must tackle each stakeholder and every step of data transfer. This model delivers an approach of where to start and how to embed further privacy analysis. However, as it is rather concerned with the flow of data, a different privacy model for the data per se is needed.

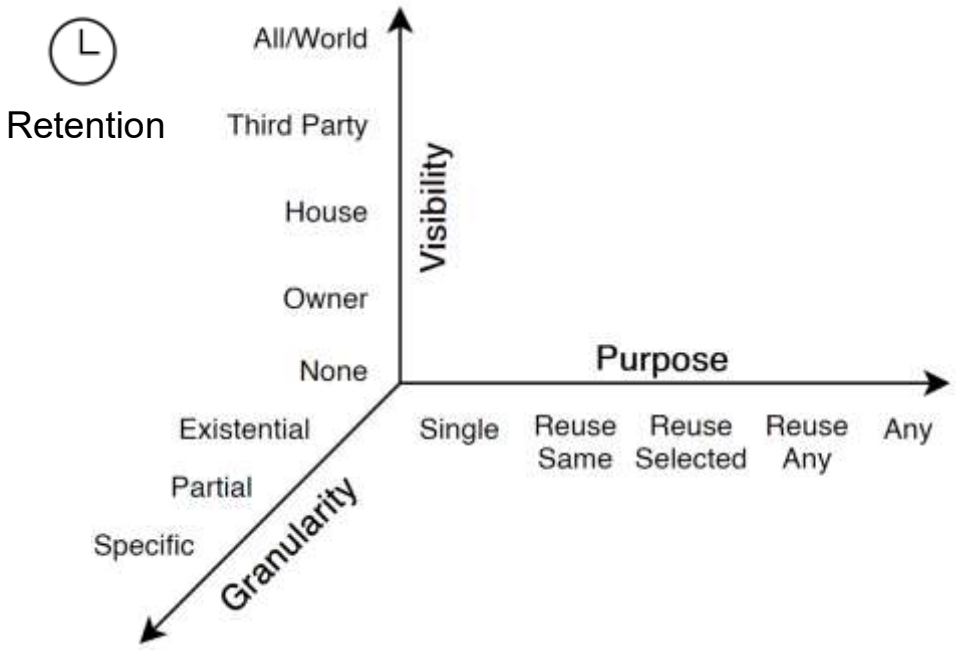


Figure 6: Key contributors to data privacy in a data repository (modified after BARKER 2009: 44).

According to BARKER et al. (2009: 53) privacy is multifaceted and therefore not easy to deal with. They make a distinction between four dimensions of privacy from the data owner perspective: "purpose, visibility, granularity, and retention" (ibid.: 44), i.e. what the data are passed on for, to whom they are visible, the data resolution, and how long they are stored or used (see fig. 6).

In practical considerations, this concept was partially taken up and, in the difficult balancing act between privacy and accuracy of the data based on the four facets of LBSM already mentioned. LÖCHNER, DUNKEL & BURGHARDT (2018: 3f) recognize that the individual facets must have different levels of resolution depending on the intended use. For example, for a rescue team in disaster management, precise coordinates and a time stamp of posts accurate to the second decide on life and death, while for journalists, a coarsely resolved location reference (e.g. city) and daily accuracy are sufficient (ibid.). With respect to fig. 6, these examples only refer to the granularity on the x-axis but would also need to have different purposes (y-axis) and visibilities (z-axis). For example, the rescue team would only need a single data use to rescue a person and visibility just for the rescue team whereas the journalist might reuse some data and wants to publish them.

For an efficient privacy evaluation, the data flow of XU et al. (2014: 1151, fig. 5) and the privacy concept of BARKER (2009: 44, fig. 6) need to be combined.

2.4.2 Geoprivacy

In this context, the so-called geoprivacy as subcategory of data privacy plays a particular role. As KESSLER & MCKENZIE (2018: 5) see it, "[i]nformation about an individual's location is substantially different from other kinds of personally identifiable information" as they allow "for a broad range of location-based inferences, such as information about their health, consumer behavior, or social status" (ibid.: 7).

Despite a high general concern about privacy matters in population (ibid.: 9), there seems to be little consciousness about geoprivacy in everyday life when sharing locations with services and apps (ibid.: 7). A recent study of CHEN et al. (2021: 27) even shows that "[…] "that there is no relationship between privacy concern […] and the number of data-sharing authorizations, confirming the puzzling data privacy paradox". According to the authors the reason is "economic benefits of sharing personal data with mini-programs" (ibid.).

Unfortunately, such a loose laissez-faire attitude is a trend that can also be observed in science. Notwithstanding broad privacy discussions, e.g. during the 2005-2012 timespan, the scientific publication of "unmasked confidential [geo-] data increased" (KOUNADI & LEITNER 2014: 140).

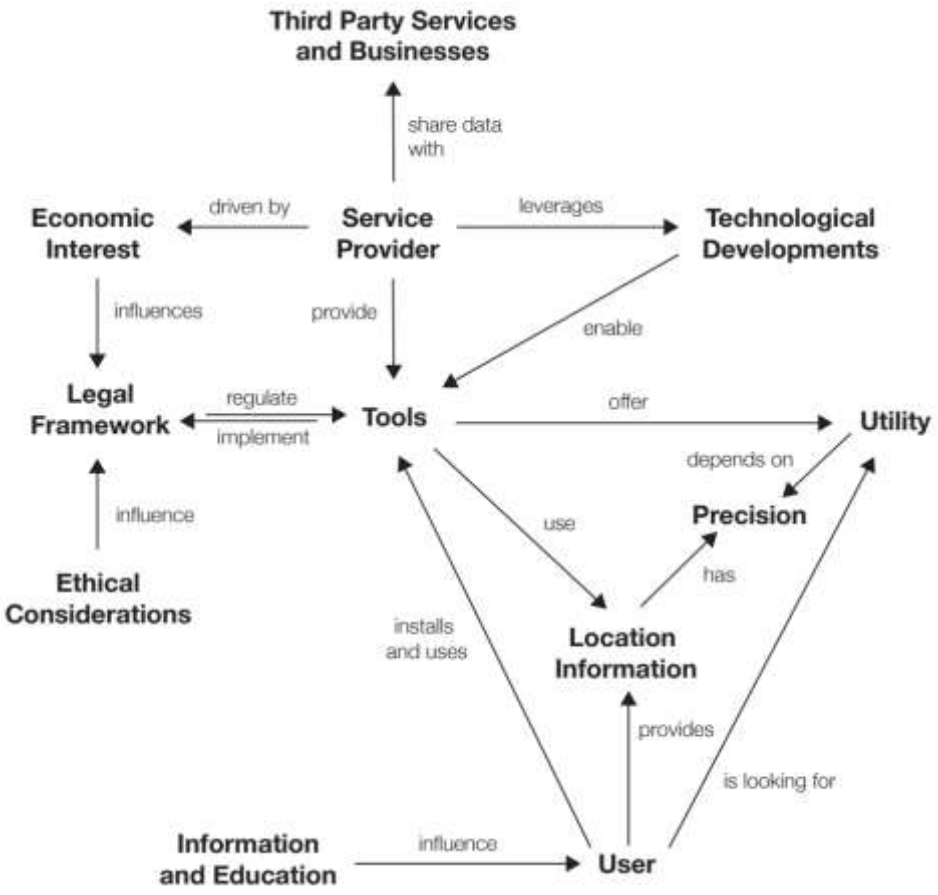


Figure 7: Tension field geoprivacy (KESSLER & MCKENZIE 2018: 16).

According to KESSLER & MCKENZIE (2018: 14ff), geoprivacy is a complex tension field between societal developments, useful services fueled by user data, legal and ethical considerations making it "difficult to tackle as a whole" (ibid.: 16, fig. 7).

MCKENZIE, JANOWICZ & SEIDL (2016: 159) show how much unintentional information can be revealed from semantic data bands. LIU (2007: 1430) argues that geoprivacy needs strong protection respectively "more privacy requirements than merely location k-anonymity" (see ch. 3.2 for k-anonymity).

However, not necessarily conflicting with the geoprivacy manifesto of KESSLER & MCKENZIE (2018: 5) but rather extending the reasonable privacy concerns, not only geodata reveal plenty of personal information but also temporal, e.g. daily, weekly or yearly distributions (MCKENZIE, JANOWICZ, & SEIDL: 166) and thematic data, e.g. what is posted about (ibid.: 167ff).

2.4.3 LBSM Big Data Privacy

The conclusion to be dawn here, is not to weigh the level of privacy risk potential of different data facets against each other, but rather to point out the importance of a high general privacy for all data and accept the complexity of the tension field (fig. 6). As big data bear a high analysis potential, they are particularly vulnerable and need a strict privacy awareness.

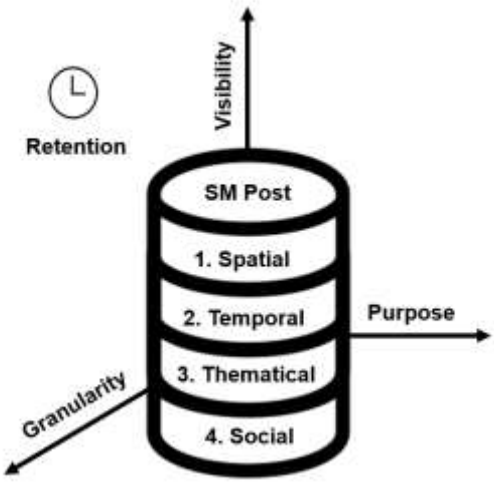


Figure 8: Post facets and privacy dimensions (modified after LÖCHNER, DUNKEL & BURGHARDT 2018: 2; BARKER 2009: 44).

Summing up this chapter, fig. 8 provides an overview of the various starting points for LBSM privacy measures, which is to be based on the four dimensions of privacy according to BARKER (2009: 44) and the four facets according to LÖCHNER, DUNKEL & BURGHARDT (2018: 2) centered around LBSM data. This model is to be understood embedded in a data flow as described by XU et al. (2014: 1151, fig. 5) and hence to be evaluated for every data stakeholder.

As working with sensitive data generally is a delicate matter it requires not only prudence but also a profound discussion of potential risks, ethical and legal implications. Unfortunately, as HEIKINHEIMO et al. (2020: 20) put it, "legislation and ethical guidelines on using publicly but passively contributed data sets in research are still immature" due to its novelty. However, the respective discussion in the concrete context of this LBSN dashboard follows in ch. 3.3.3.

How privacy can be tackled in the LBSN dashboard is presented in ch. 3.2 and evaluated based on the privacy model in ch. 5.

3 Methodology

It should be noted that the focus of this thesis is to develop a LBSN dashboard, which is intended to apply the quantitative methods defined here to certain research questions or in the context of policy-making. The concrete, actual application of the methods is therefore not the main goal here, but is only carried out in the case study as an example for the Bonn area for illustration purposes.

This thesis is also application-oriented. Methodically, due to the big data character of the LBSM data, quantitative methods are used, which consist in particular forms of data retrieval and analysis. The application of qualitative methods is not treated in this thesis but is proposed in the outlook (ch. 4.6).

This chapter is divided into a general classification of quantitative methods of social research, a more practical HLL introduction for LBSM big data and finally a detailed description of the developed dashboard.

3.1 Quantitative Research

Before a more detailed methodological discussion can take place, it must first be clarified what exactly is being analyzed and how. In ch. 2.2 it was worked out that LBSM can be divided into four facets. Of these four facets, the thematic facet is analyzed through the textual information expressed in terms, in the context of the social, spatial, and temporal facets.

Here, an overview is given of how textual analysis can be methodologically situated.

When researchers speak of quantitative analysis, it does not yet indicate what exactly is being studied and how. BERNARD (2013: 393) calls for a precise definition of what is studied and how. For this purpose, he has set up a simple matrix that illustrates the difference between data character and the character of analysis (ibid.; ibid. 1996: 10). It is adapted and extended here for the use case of this thesis.

Table 3: Qualitative-quantitative data and analysis (modified after BERNARD 2013: 393; 1996: 10). The dashboard focus is marked in blue (cell c & d), the dashboard purpose in green (cell b).

		Data	
		Qualitative	**Quantitative**
Analysis	**Qualitative**	(a) Interpretative text studies. Hermeneutics, Grounded Theory	(b) Search for and presentation of meaning in results of quantitative processing
	Quantitative	(c) Turning words into numbers. Classic Content Analysis, Word Counts, Free Lists, Pile Sorts, etc.	(d) Statistical and mathematical analysis of numeric data

Tab. 3 refers to LBSM and therefore primarily to text data. LBSM data initially represent qualitative data that can be analyzed either qualitatively or quantitatively.

Starting from the researcher's perspective with raw data, i.e. the full LBSM data with all information, a qualitative analysis could be carried out and, for example, in the course of grounded theory (see GRUBER & HOLSTEIN 2014), search for connections of meaning in individual posts (tab. 3, cell a).

However, since this work is application-oriented and assumes a complete workflow, within which the raw data are not stored in the course of privacy (explained in detail in ch. 3.3.4), this option is omitted for practical purposes.

The context is different with the quantitative analysis of qualitative data (tab. 3, cell c). This type of analysis is mainly concerned with quantifying the qualitative data such as words and hashtags. The simplest form of quantification is, for example, word counts, i.e. how often a word was used at a certain location and how often in a certain period of time. Such word counts – in different variations – represent the essential approach of how the data are made accessible to laymen by the dashboard developed here. Word counts are purely descriptive data, but they contain a high potential for further analysis as is shown later.

The results can then be displayed and interpreted immediately either as a frequency list, bar charts or on a map using various methods (e.g. markers, heat maps, bins), which, according to BERNARD (2013: 393; 1996: 10), corresponds to the qualitative analysis of quantitative data (tab. 3, cell b).

Alternatively, the results could also be correlated or intersected with other data that can be added to the dashboard, which corresponds to the quantitative analysis of quantitative data (tab. 3, cell d). For example, in the simplest case, a time series of word counts for the term "restaurant" on LBSM could be correlated with the number of actual existing restaurants.

Afterwards, it is always crucial to interpret the results (tab. 3, cell b) in order not to come to the wrong conclusions.

Simply speaking, either the thematic component of LBSM can be transformed into numbers by word counts (tab. 3, cell a) and then visualized, analyzed and interpreted (tab. 3, cell b) or these two steps can be interspersed with the further analysis by quantitative data.

This method section mainly focuses on quantitative analysis of qualitative data (tab. 3, cell c) and subsequent presentation as well as interpretation. The extent to which LBSM big data play a special role in this is now clarified.

3.2 HyperLogLog and Location-Based Social Media Big Data

Based on a cross-topic literature review, LBSM big data are defined here following DE MAURO, GRECO & GRIMALDI (2015: 103) as LBSM data "characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value [sic]."

ILIEVA & MCPHEARSON (2018: 555) point out the high LBSM big data potential for "large-scale social–ecological analyses" but conclude that "comprehensive accounts on the use of geolocated big data from social media for sustainable city planning are still rare" (ibid. 553).

Numerous big data leaks have happened in the past in the SM context and might have contributed to the rare LBSM big data application in practice. The problem with the data is, that it can be easily

duplicated and true to the motto "The internet never forgets" and contrary to Article 17 of the GDPR (2016: art. 17), the "Right to be forgotten"[7] can be disseminated uncontrollably.

An LBSN dashboard must therefore meet particularly strict privacy requirements, especially if it is to be publicly accessible. On the one hand for legal (ch. 3.3.3.1) and on the other hand for ethical (ch. 3.3.3.2) reasons. In addition, with conventional data processing of raw data, the processing time increases proportionally with the amount of data and consumes a corresponding amount of disk space.

Compared to the myriad of conventional LBSM research papers (ch. 2.2), few papers have already approached this problem with a technical solution in the SM context. In this context, a proportionally large number of works are concerned with k-anonymity. With k-anonymity, an attempt is made to describe whether personal data can be traced back to a person (GONG, SUN & XIE 2010: 367). The main idea here is to aggregate or generalize the data to such an extent that sufficient, i.e. at least $k-1$ elements with a minimum number n with the same characteristics are present in the total set, thereby excluding unambiguous traceability to a person (SAMARATI & SWEENEY 1998: 5ff). In the SM context, k-anonymity has been intensively researched with a geospatial focus (e.g. WANG et al. 2018; ZHAO et al. 2018; NIU et al. 2014; LIU et al. 2013; GEDIK & LIU 2004).

Another approach to create privacy-aware data models in the LBSM context lies in probabilistic data structures such as bloom filters differential privacy (e.g. WANG & SINNOTT 2017), (e.g. TANG, REN & ZHANG 2018; ASADI & LIN 2013) or HLL (e.g. LÖCHNER et al. 2020). While these mathematical approaches are by no means new, they are only explored by a tiny community in the LBSM context worldwide.

In their benchmark implementation, DUNKEL, LÖCHNER & BURGHARDT (2018: 3f) compare conceptual weaknesses of the above-mentioned methods and find that the HLL method has certain advantages in comparison and especially in practice. The latter is discussed in sufficient depth below, whereas no detailed comparison of all procedures can be made here. Since this thesis is fundamentally based on HLL as a privacy approach, it is presented here with regard to the target group of this thesis from an application-oriented perspective[8].

3.2.1 Count-Distinct Problem

The count-distinct problem deals with how cardinalities can be calculated as efficiently and accurately as possible, i.e. how distinct elements can be identified from a set. In the context of LBSM, this is an elementary problem that can already provide important information e.g. on visit frequencies, utilization rates, or serve as site analysis. A simple example would be the information of how many users were at

[7] For a more detailed explanation see POLITOU, ALEPIS & PATSAKIS (2018).
[8] For mathematical details instead, see the original paper by FLAJOLET et al. (2007).

a certain place in a certain period of time or how many posts with certain terms were posted in a certain period of time.

With original data, this problem could be solved by simple SQL filters, for example, if the posts are filtered by time period, location and any terms they contain. The filter leads to a set of posts, within which a unique identifier (UID), such as a user ID or nickname (available on Facebook, Instagram, Twitter, etc.), but also an email address, telephone number, etc., can be used to determine how many unique elements are in this particular set using a classic count-distinct algorithm. If, for example, one is interested in how many different users regularly post in one place, each user may consequently only be counted once, regardless of how much each user posts.

This approach is problematic for two reasons. On the one hand, the computational effort and storage requirements increase proportionally to the number of posts, which, in view of the big data character of LBSM data, would lead to cost-intensive computations of powerful central processing units (CPUs) and constantly expanding storage capacities. If such a dashboard is to be hosted by a municipality or even by individuals, this would already be a criterion for exclusion, depending on the available financial means.

On the other hand, different elements in a subset can only be identified if UIDs are permanently stored and extended with all metadata like timestamp (date, time), location (coordinate, location ID o.s.) and so on. With regard to fig. 5, this would mean that all original data would have to be delivered from one data stakeholder to the next, i.e. through the instances of data provider, collector and miner. Only in the last step, from transmitter to decision maker the counts would be sufficient. This flow of data would make the original data vulnerable by permanently storing it in multiple instances and forwarding them. This procedure is thus ruled out for a potentially public dashboard.

At this point, HLL comes into play as it is not only able to reduce storage and computational requirements to a minimum, but also to contribute significantly to privacy through its probabilistic nature when used correctly, as the original data set cannot be restored from the HLL format. Furthermore, it does not allow to store the raw data at any time, but instead processes it directly in-memory.

3.2.2 Introduction

The HLL algorithm was developed by a group of French researchers around the mathematician FLAJOLET et al. (see 2007; FLAJOLET & MARTIN 1985). It provides a solution to the count-distinct problem with a special focus on low memory and computational resources by transforming raw into probabilistic data.

The core idea of HLL can be illustrated by a Laplace coin toss experiment, given a coin with a “1” (heads) on one side and a “0” (tails) on the other, where both sides are equally likely to occur (50%).

The probability in an experiment *e* where the coin is tossed five times to get "0" (or tails) the first four times in a row is improbable. More precisely, this probability is

$$\frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} = \left(\frac{1}{2}\right)^4 = \frac{1}{16} = 0.0625 = 6.25\%.$$

However, if this experiment is carried out often enough, the probability increases proportionally, since each run of five tosses has the same probability.

Table 4: Fictive Laplace experiment with 14 runs.

Run Number	Coin Series	Leading Zeros
1	**0** 1 0 1 1	1
2	1 1 1 0 1	0
3	**0 0** 1 0 0	2
4	1 0 1 0 1	0
…	…	…
14	**0 0 0 0** 1	4

Tab. 4 shows a fictive experiment with 14 fictive runs, each tossed five times. The resulting coin series indicates whether the coin displayed a zero ore a one. Here, the leading zeros (LZ), i.e. how many times in a row, the first tosses displayed a zero, are marked in bold. In this fictive experiment, on the 14th run of five tosses, the first four tosses are zeros, which is the longest series of leading zeros (LSLZ).

The idea of HLL is exactly the opposite, i.e. to derive from the LSLZ how many runs there were. Assuming that the LSLZ for another fictive experiment is five and that one does not have any other information, it can be deduced that the probability is

$$\left(\frac{1}{2}\right)^5 = \frac{1}{32} = 0.03125 = 3.125\%.$$

On average, five LZs occur in only one out of 32 cases and thus one would guess 32 coin tossing runs for this experiment.

Table 5: Probability estimation based on leading zeros.

Binary sample event series e with $Z(e)$ leading zeros	Longest series of leading zeros $Z(e)$	Leading zeros series probability $P(Z(e)) = (½)^{Z(e)}$	Estimated number of runs $N(e)_{est} = 2^{Z(e)}$
1 0 1 1 0 0 0 …	0	$(½)^0 = 100\%$	$2^0 = 1$
0 1 0 0 1 0 1 …	1	$(½)^1 = 50\%$	$2^1 = 2$
0 0 1 0 1 0 0 …	2	$(½)^2 = 25\%$	$2^2 = 4$
0 0 0 1 1 0 0 …	3	$(½)^3 = 12.5\%$	$2^3 = 8$
0 0 0 0 1 1 0 …	4	$(½)^4 = 6.25\%$	$2^4 = 16$
0 0 0 0 0 1 0 …	5	$(½)^5 = 3.125\%$	$2^5 = 32$
0 0 0 0 0 0 0 1 …	6	$(½)^6 = 1.5625\%$	$2^6 = 64$
0 0 0 0 0 0 0 …	7	$(½)^7 = 0.78125\%$	$2^7 = 128$

Tab. 5 shows the estimation process which is independent from the real number of runs $N(e)_{real}$.

In practice, however, the binary strings do not represent random coin tosses, but actual information such as UIDs. Assuming that some hypothetical user UIDs are randomly assigned, they only need to be converted from UTF-8 i.e. "standard" text notation to "computer-readable" binary notation. There are simple methods to convert between them (see WECKMÜLLER 2021b: section "misc"). For example, the binary representation of the UID "U123" is "10101011100011100101100 11".

This strongly simplified HLL version could already be used in reality with certain restrictions. For a fictive set of 10 000 UIDs in binary format, the LSLZ would be identified, which can be done either sequentially or parallelly. Since it is sufficient to store only the LSLZ for the whole HLL set, correspondingly little memory is needed, which depends only on the highest LSLZ considered.

Now further elements can be added very easily by simply identifying the LSLZ for each further string to be added. If it is larger than for the entire HLL set, the LSLZ would be adjusted accordingly. If it is smaller, the LSLZ for the HLL set remains the same. This procedure is called HLL union and is explained in ch. 3.2.4. Similarly, entire HLL sets can be combined if, for example, the question is how many unique users were at two different locations assuming a separate set for each location.

Individual elements cannot be removed retrospectively because there is no clear assignment due to the fact that only the first digits are considered. Accordingly, it is also not possible to check whether a certain element e was read into the set[9].

3.2.3 Theory

The method described so far is a highly simplified version of HLL, which has a high error rate. FLAJOLET et al. (2007) extend it to obtain more accurate estimates. By dividing the binary string into so-called registers (ibid.: 129), i.e. sections, an accuracy of the estimate of approx. 2% and a memory consumption of only 1.5 kilobytes can be obtained for a set size of > 1 billion (1 000 000 000) elements (ibid.: 127).

For example, a binary string can be divided into predefined registers of n bits each (tab. 6, left). The LSLZ is then only counted for the respective register. Alternatively, the first n bits can be used as random binary register number and the rest for the register LSLZ (see tab. 6, right). Depending on how many bits are to serve as register number, one gets a certain number of registers in binary logic. For example, if the first two bits are selected, four different registers are used: "00", "01", "10" and "11", which in decimal notation correspond to "0", "1", "2" and "3".

[9] For exactly this purpose, to check whether an element occurs in a set, the Bloom Filter was developed, which is based on similar principles and already has been applied to SM context (e.g. ASADI & LIN 2013).

Table 6: HyperLogLog register methods comparison.

	Fixed register width of four bits				Register number based on first two bits
Sample string	10001101001001011				
String division	1000	1101	**00**10	1011	**10** **00**1101001001011
Register number	0	1	2	3	Binary: 10 → Decimal: 2
Register LSLZ	0	0	2	0	2

For cardinality estimation, FLAJOLET et al. (ibid.: 129) calculate the harmonic mean of the probability estimates of all registers.

3.2.4 Operators

Main HLL operators are so-called unions (DUNKEL, LÖCHNER & BURGHARDT 2020: 8) which are loss-less (GARCIA-RECUERO 2021: 2) and intersections (ibid.: 12) producing rather high error rates (ibid.: 2) which is shown in the case study (ch. 4.4.2 & ch. 4.4.4).

Unions and intersections follow the inclusion-exclusion-principle (ibid.: 4; ERTL 2017: 2, 34). For two elements A, B the following equations (eq.) apply:

$$|A \cup B| = |A| + |B| - |A \cap B| \quad (1)$$

$$|A \cap B| = |A| + |B| - |A \cup B| \quad (2)$$

$$|B \setminus A| = |A \cup B| - |A| \quad (3)$$

$$|A \setminus B| = |A \cup B| - |B| \quad (4)$$

(DUNKEL, LÖCHNER & BURGHARDT 2020: 12; ERTL 2017: 34; ANDREESCU & FENG 2004: 117ff).

Eq. 1 represents the first of two important HLL equations, the standard HLL union or put simply, the number of unique elements in a merged set of A and B. Eq. 2 is the second most important equation, the standard HLL intersection or in other terms the number of unique elements occurring simultaneously in both A and B.

Further eq. 3 and 4 deliver the result of a HLL Union of A and B less unique elements of A or respectively B.

For three elements A, B, C the HLL intersection formula, i.e. the number of unique elements occurring simultaneously in all three HLL sets A, B and C, can be rewritten to

$$|A \cap B \cap C| = |A \cup B \cup C| - |A| - |B| - |C| + |A \cap B| + |A \cap C| + |B \cap C| \quad (5)$$

(ibid.).

These logical operations are quite simple but lead to the previously mentioned handy feature of HLL, that different HLL-Sets can easily be united. However, certain restrictions to HLL intersections apply, shrinking their utility to a minimum when HLL set sizes differ strongly (DASGUPTA et al. 2015: 20).

3.2.5 Hashing

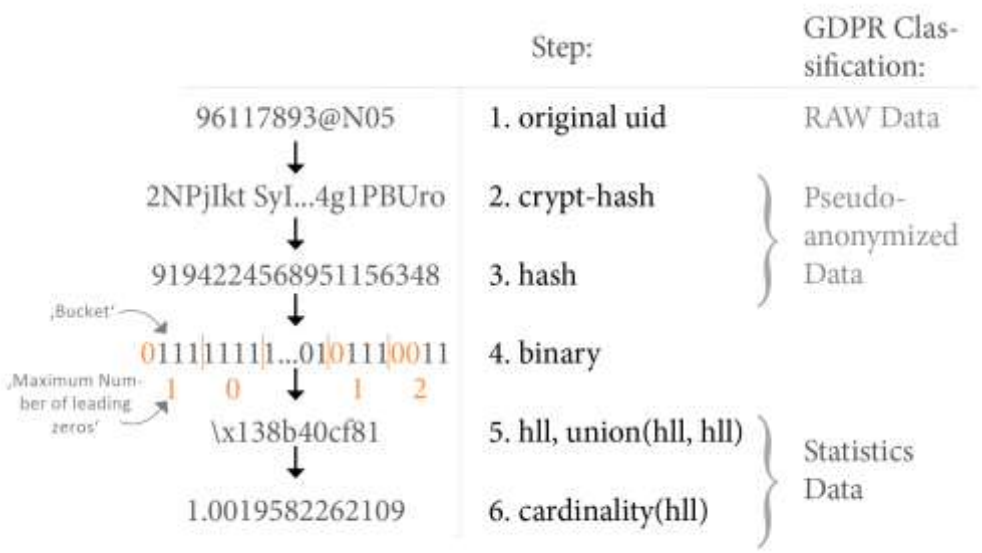
In real-life SM the distribution of UIDs is not random. Users might have certain preferences in user names, numbers etc. but also a company could choose to add a certain prefix to an UID. If i.e. the UIDs of a company always start with the prefix "User-", it could hypothetically translate to eight LZ right in the beginning of any bit string. This scenario would interfere with the necessary HLL assumption of a uniform distribution of bit series.

Hash-functions can help with this. A hash function is a function that maps a string with an arbitrary number of characters to another string with a fixed number of characters without collision (DAMGÅRD 1989: 416). A UID, e.g. "User-123456789" could thus be mapped to a string of e.g. six characters and result to "x9jQwR". This in turn can be represented binary, i.e. only with ones and zeros. In addition, there are certain specifics, such as the greatest possible difference between the hashes of similar input IDs. For example, the hash of "User-123456789" must be very different from the hash of "User-123456788" and should not be too similar, e.g. "x9jQwT". For a real-life implementation a uniform hash function is vital. Only if the hash function is capable of distributing bits "over all m registers according to a multinomial distribution with equal probabilities" (ERTL 2017: 2), it is suited for usage in combination with HLL[10].

Additionally, DUNKEL, LÖCHNER & BURGHARDT (2020: 8) use cryptographic hashing as proposed by DESFONTAINES, LOCHBIHLER, & BASIN (2019: 27f) as an additional step to make HLL-Sets more secure and private (DUNKEL, LÖCHNER & BURGHARDT 2020: 7). The step of hashing, encrypting and transforming to binary representation is also referred to as "Sketching" and the result as HLL-Sketches (ERTL 2017: 1).

[10] For an evaluation of existing hashing algorithms suited for HLL see DAHLGAARD, KNUDSEN & THORUP (2017). The PostgreSQL-HLL implementation (CITUS & CONTRIBUTORS 2021) uses MurmurHash 3 (see APPLEBY 2021).

Table 7: Transformation steps applied to a single character string, such as a user ID, for generating a HyperLogLog set, and the final estimation of cardinality (DUNKEL 2020: 7; DUNKEL, LÖCHNER & BURGHARD 2020: 7).



As indicated in tab. 7, the original UID can firstly be hashed cryptographically. Afterwards it is hashed for an even distribution, converted to binary and read into an HLL-Set. As DUNKEL (2020: 7) states, different GDPR classifications apply to different steps. Starting with raw data, they can be turned into pseudo-anonymized data by hashing. HLL sets fall under definition of statistics data and are hence much less restricted (ibid.). For example, the "Right to be forgotten" (GDPR 2016: art. 17) as mentioned earlier does not apply here.

3.2.6 HyperLogLog Application

HLL is intended for cardinality estimation through HLL sets describing a certain attribute. In the examples mentioned here, this attribute was always the UID. So, the cardinality to be determined in all examples so far refers to the number of unique users.

Just as well, it can also refer to the post ID (which should probably already be distinct) or to the date. For the latter, it could be determined, for example, for how many different days in the year there were posts. Since this would not provide much insight for large datasets, because people probably post on all days of the year, a combination of users and days, so-called userdays, is formed (DUNKEL, LÖCHNER & BURGHARDT 2020: 5). Userdays are days on which a user has posted something without regard to quantity.

Such a combination can be easily created when reading the original data into a new HLL set by concatenating the user ID and the date in a uniform format. In tab. 7, for example, an "@" character is used between the first and second part (see DUNKEL, LÖCHNER & BURGHARDT 2020: 7), which for the example user ID "U123" for the date 07/01/2020 (format DD/MM/YYYY) could look like: "U123@01072020". With the help of these basic metrics, data user count, post count and user days

very exact cardinalities can be determined. The question is, for what exactly, i.e. for which bases these metrics are determined.

3.2.7 Social Media Facets

For this purpose, certain subsets can be created in advance. For example, if one always wanted to monitor all data of the previous day, these would be read into a separate HLL set. Retroactively, these can have not only arbitrary temporal filters, but also spatial filters by filtering the original data before creating the set. In this way, with respect to the four SM facets (ch. 2.2, fig. 4), the first two (temporal, spatial) are already covered.

The social facet can already be roughly covered by the selection of the LBSN, since each LBSN has a different user base (ch. 3.3.1). Furthermore, simple filters (e.g. only male users over 30) can be applied or read in like user days in combination with other attributes as long as the data have such a high resolution. For example, even very abstract base combinations for the skewness of a distribution like "Genderdays" (m/w/n@date) or "Userages" (UID@age) could be formed. The latter combination would e.g. be an indicator for the degree of age mixing at certain locations. These combinations have not been subject to research and should be investigated for further usage as their interpretation is vital for decision-making.

Such base combinations can also be generated with locations, coordinates or geohashes. However, the use of coordinates only makes sense if they occur more than once. On Flickr, for example, individual coordinates are often stored for posts, whereas on Instagram only the location IDs are stored causing a coarser resolution. Still, these location IDs can be used for combinations.

In contrast to the other facets, the procedure for the thematic facet is less intuitive. The key here is to break down the actual content of the posts into its "atomic" components as proposed by DUNKEL et al. (2021: n.p.) by their "Location Based Social Network (LBSN) structure" i.e. a "common language independent, privacy-aware and cross-network social-media data scheme" that is referred to as LBSN-Structure in the following.

With this data scheme, the connection between UID and post content are dissolved, as this link is highly problematic from a privacy perspective (DUNKEL, LÖCHNER & BURGHARDT 2020: 15). This procedure not only generates a large part of privacy, but is also the foundation for HLL.

Table 8: Fictive sample social media post with metadata.

SunLover22 • 30/07/2021, 02:15 P.M. • Capo di Conca, SA, Italy
Today enjoying the *wonderful* ⚙ on the ⛱ 🌊 🌴 #costaamalfi with @lovingsister #siblings # sun #beach #summer 😊
[Photo with accessibility caption: "This image could contain 2 persons on the beach"]

According to this procedure, posts (tab. 8) can be divided into their atomic components consisting of metadata such as UID, timestamp and location, the post content consisting of text, emojis, hashtags, user tags and a photo as well as photo metadata amongst which for the scope of this thesis only the accessibility caption is of interest, as e.g. photo dimensions, format and used camera/phone model do not matter here (table 9).

Table 9: A fictive social media post split into its atomic components.

Category	Post component	Sample
Post Metadata	UID	SunLover22
	Timestamp	30/07/2021, 02:15 P.M
	Location	Capo di Conca, SA, Italy
Post Content	Terms	"Today", "enjoying", "the", "wonderful", "on", "the" "with"
	Emojis	"⚙", "⛱", "🌊", "🌴", "😊"
	Hashtags	"#costaamalfi", "#siblings", "#sun", "beach", "summer"
	User Tags	"@lovingsister"
	Photo	[A photo of two persons on the beach]
Photo Metadata	Accessibility Caption	"2 persons", "on the beach"
	…	…

The accessibility caption for the visually impaired can be activated in order to be read by text-to-speech-plugins. The caption is automatically generated by the respective LBSN mostly based on Machine Learning (ML) algorithms[11].

[11] There is no scientific study yet, that has made use of HLL sets for accessibility captions. However, this seems to be a topic with high research potential, as the captions provide standardized information that can be used without having to process the photos themselves and putting them at risk. The caption can just be treated like regular text with the only exception that certain information is only valuable in combination like "2 persons". Further research is needed in this area.

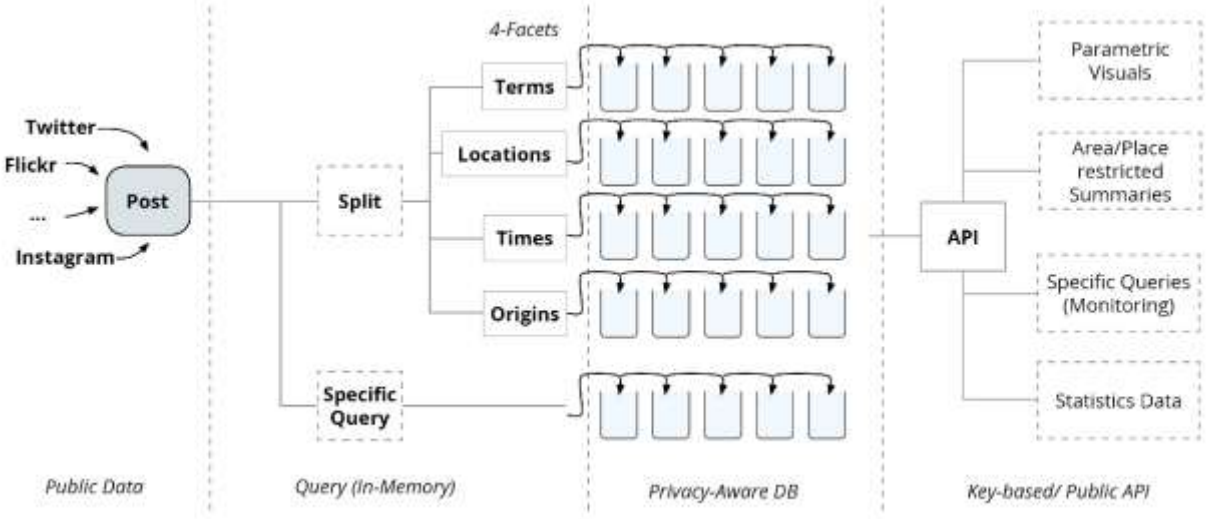


Figure 9: HLL data workflow for splitting social media posts in its atomic components (DUNKEL 2020: 19).

All this information is thus separated from each other, i.e. each atomic piece of information is read into the respective HLL set (fig. 9). If there is no HLL set for the respective word, a new one is created leading to a many different HLL sets, all well independent from each other. For example, the HLL set for the hashtag "#costaamalfi" would now have a cardinality of approximately 1 if no other post used the same hashtag before.

Considering the wide field of Natural Language Processing (NLP), plain text information processing could be taken further with methods such as stemming and lemmatization but are not discussed here[12].

3.2.8 Privacy

As the main idea of this thesis is to use HLL in combination with other methods in order to preserve privacy as far as possible but at the same time to ensure a certain scope for analysis, it must be evaluated on the one hand how HLL itself contributes to privacy and on the other hand which privacy risks arise in the context of a real-life application. The former is clarified in this chapter, the latter after treating the case study in ch. 5.

DESFONTAINES, LOCHBIHLER & BASIN (2019) provide an excellent overview of potential privacy risks with cardinality estimators such as HLL. They advocate caution in general, arguing "that cardinality estimators should be considered as sensitive as raw data" (ibid.: 41). They even go so far as to claim in their paper title that "Cardinality Estimators do not Preserve Privacy" (ibid.: 41). This statement, however, refers to a public database or an attacker who has full, i.e. writing access to the entire database and the hashing function. In this case, HLL is indeed not privacy-preserving, since an attacker can generate the corresponding UID hash for his target, read it into all existing HLL sets, and check whether

[12] For a detailed overview of NLP methods see EISENSTEIN (2019).

the counter changes (ibid. 2019: 33), which is the worst case. If so, the target was guaranteed not to be in the HLL set and thus not to be in a particular location, for example. When tracking a user "across locations and time periods" in order to derive more information, this is also referred to as "intersection attack" (ibid.). LÖCHNER, DUNKEL & BURGHARDT (2018: 9ff) describe this scenario in detail with two fictive examples in order to understand the worst case and evaluate its probability.

For the opposite case, however, if an attacker reads a target UID into the HLL-DB and the counter does not change, no information can be gained, since no probability can be derived about whether the target's hash has already been read in or all registers already count the same or an identical number of LZs, which explains the effect of "hiding in the crowd" (FEYISETAN et al. 2019: 2, as cited in DUNKEL, LÖCHNER & BURGHARDT 2020: 13) similar to k-anonymity. In other words: an attacker must be lucky to hit a target hash which would provoke a change, otherwise no valuable information can be derived.

However, the smaller the dataset the higher the probability that the counter changes and vice vera (DESFONTAINES, LOCHBIHLER & BASIN 2019: 39). This behavior characterizes the tradeoff between accuracy and privacy where certain limits should be considered. At this point the conclusion to be drawn from DESFONTAINES, LOCHBIHLER & BASIN (2019) is simple: a publicly accessible, unencrypted HLL-Set is vulnerable and hence not privacy-preserving.

Instead, they propose certain risk mitigation strategies:

- Conventional data protection such as "encryption, access controls, auditing of manual accesses" (ibid.: 15)
- Cryptographic hashing functions (ibid.:15)
- Minimum cardinality (ibid.: 7) which is similar to a minimum HLL set size
- Salt-keys for different HLL sets (preventing HLL-union possible) (ibid.: 15)
- Homomorphic encryption (high computation cost) (ibid.: 16)

Additionally, REVIRIEGO & TING (2020: 4) propose keeping two HLL-Sets parallelly, one salted the other without in order to monitor possible manipulation attempts by the difference of the two sets.

In a real-life application when only some these measures implemented, external intersection attacks become very unlikely to retrieve valuable information in comparison to the attacking effort and anterior knowledge required.

Nevertheless, a future field of research should be an optional addition of noise to the HLL sets in order to fulfill the definition of differential privacy. It should be investigated, to what extent noise is increasing the error rate and where an appropriate balance of necessary noise for differential privacy on the one hand and usability as well as a low error rate.

However, a crucial point in the context of this thesis is the fact that all the data to be read into the database are publicly available. Public means without having an account or special access rights on the respective LBSN but freely accessible for everyone online. Considering this important requirement, discussing possible complex attacking scenarios seem almost pointless as any potential attacker would always search for the original information in the respective LBSN as it is easier, more abundant and faster. If an attacker already knows the UID there is simply no point in trying to get access to the hashing function, finding out the right salt-key as well as encryption function and testing HLL sets for the UID hash. Only in the special case where the original post was deleted meanwhile such an attacking effort would make sense as data cannot be deleted from HLL sets. Even in this worst-case scenario, the information scope is highly limited if all privacy measures, such as whitelists (ch. 3.3.4) are applied.

3.2.9 HyperLogLog Conclusion

HLL is a fast, low on memory and quite accurate. The algorithm itself does not preserve privacy per se, as in some cases, it can be identified whether an element occurs in a HLL set or not whereas the original dataset cannot be restored.

While HLL can only be called privacy-aware, a HLL infrastructure consisting of a proper cryptographic hashing function, salt-keys, access control, minimum set sizes and whitelists it becomes very close to privacy-preserving. If an attacker might overcome all of these measures, it is sometimes possible to verify whether an element is in a set or not which represents the worst case. Due to this fact, following DUNKEL, LÖCHNER & BURGHARDT (2020: 2) within this thesis the dashboard infrastructure is called "privacy-aware" instead of "privacy-preserving".

With the method of separating posts into its atomic pieces, HLL supports privacy further and hence stands on a solid ground in the balancing act between privacy and utility of the data. As all of the data used in this thesis are publicly available, on average there is a low incentive for attackers to get access to the HLL infrastructure. However, further research is needed to increase the level of privacy.

The HLL method in combination with LBSN-Structure described in this chapter is fully implemented in a PostgreSQL-based Docker container by DUNKEL & LÖCHNER (2021a) and referred to as HLL-DB. Fig. 10 sums up the data workflow and how LBSN-Structure is applied with HLL-DB.

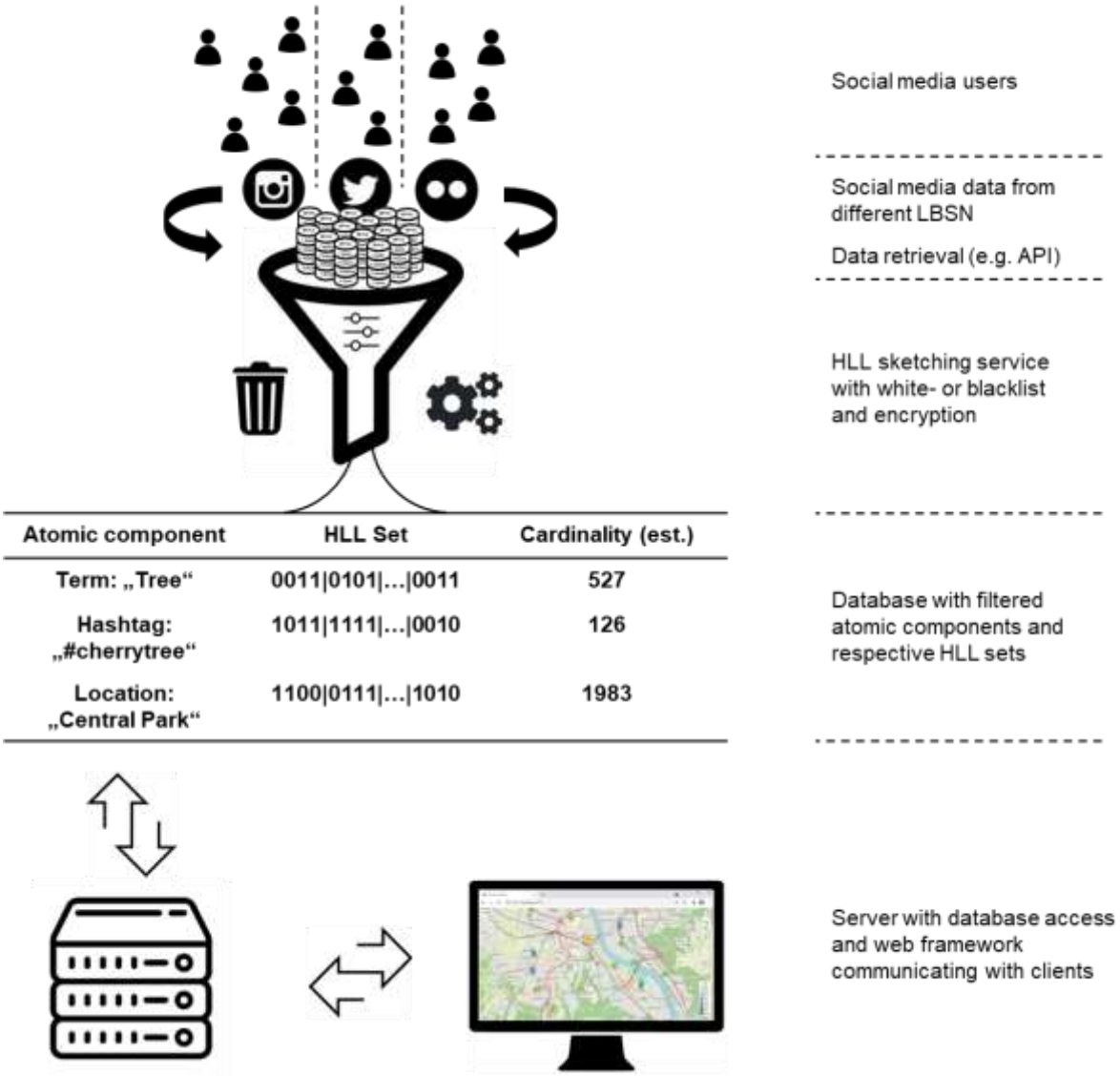


Figure 10: Dashboard data processing pipeline.

3.3 Application: The Dashboard

This chapter is dedicated to describe the working prototypical LBSN dashboard developed in this thesis. It already has a large feature set (ch. 3.3.5) but is still to be considered a prototype, i.e. most of the functions are already developed while others are still to be implemented.

For all of the following screenshots, common cartographic guidelines are neglected in favor of a more practical illustration of the actual appearance of the dashboard. By default, all maps are oriented to the north and have a scale and corresponding legend in the dashboard. These do not appear on the screenshots in favor of better formatting. More screenshots and an introductory video are available in the supplementary GitHub repository (see WECKMÜLLER 2021b).

The resolution and the size of the OpenStreetMap background map as well as its labeling is bound to the resolution and the zoom level. Since the zoom level is always chosen in favor of appearance, the following screenshots may show different OSM label sizes.

3.3.1 Data Sources and Data Mining

There is a growing and ever-changing range of LBSN. In this thesis, data from the LBSN Instagram, Twitter and Flickr are used for reasons to be clarified below. Each LBSN as well as the respective access to the data are briefly presented.

3.3.1.1 Instagram

According to latest official internal statistics of FACEBOOK (2021: n.p., as of 05/2021), Instagram has a base of >500 million daily active users (DAU) (as of 09/2017) or >1 billion monthly active users (as of 06/2018). Even though the main feature of Instagram is posting pictures which is not considered here, the platform offers the highest potential for a LBSN dashboard, as a relatively high percentage of the posts are geotagged with a so-called "Facebook location" having a certain description and a coordinate. Unfortunately, as Facebook seldomly reveals official and precise statistics the geotag quota cannot be quantified further. However, there are certain clues for a probably high usage of geotags.

1) Empirically, FIALLOS et al. (2018: 2, 4) point out, that their sample of more than 900 000 posts contained around 41% of geotags and BOY & UITERMARK (2015: 9) estimate, that an overall sample for Amsterdam might contain roughly around 21-25% of geotagged posts.
2) Business strategists and SM companies highly advocate for geotagging as it "can expand the reach of your posts" (SENDIBLE 2021: n.p.). Hence, an increasing tendency can be expected for driving sales and increasing audience.
3) While datamining, only for the area of Bonn roughly <12 000 locations (very likely more available and growing) could be retrieved from Instagram.

Facebook locations cannot be created through the Instagram app directly but rather need to be added using Facebook. Everyone can create locations at any point, either regular locations (see WECKMÜLLER 2021b: section "misc") or business locations (see LIM et al. 2016). This degree of freedom, on the one hand offers a huge analysis potential for VGI as the users directly create a location where they want their posts to be associated with. On the other hand, it leads to a massive degree of confusion in the data.

Table 10: Problems of user-created locations.

	Short description	Long description	Example
1	Wrong locations	A location has with wrong coordinates	Locations with coordinates of Bonn but name "Köln" (City of Cologne)
2	Location Duplicates	The same location appears multiple times	Multiple locations for Hofgarten (UGS in Bonn)
3	Outdated locations	A location to a spot not existing anymore	A bar or restaurant recently closed due to Corona[13]
4	Periodic locations	A periodically existing location	"Weihnachtsmarkt" (Christmas Market Bonn)
5	Fake locations	A location created for the sole purpose of trolling or misleading	"Atlantis Ocean Bar" (see WECKMÜLLER 2021b: section "misc")

For example, when mining data for Bonn, there is a coordinate with different location IDs representing the entire city of Bonn (and other city events such as "Rhein in Flammen" or the Christmas market) with coordinates right on the "Martinsplatz" in front of the "Bonner Münster" (central cathedral). Just a few meters next to it, the respective location ID referring to the cathedral itself can be found.

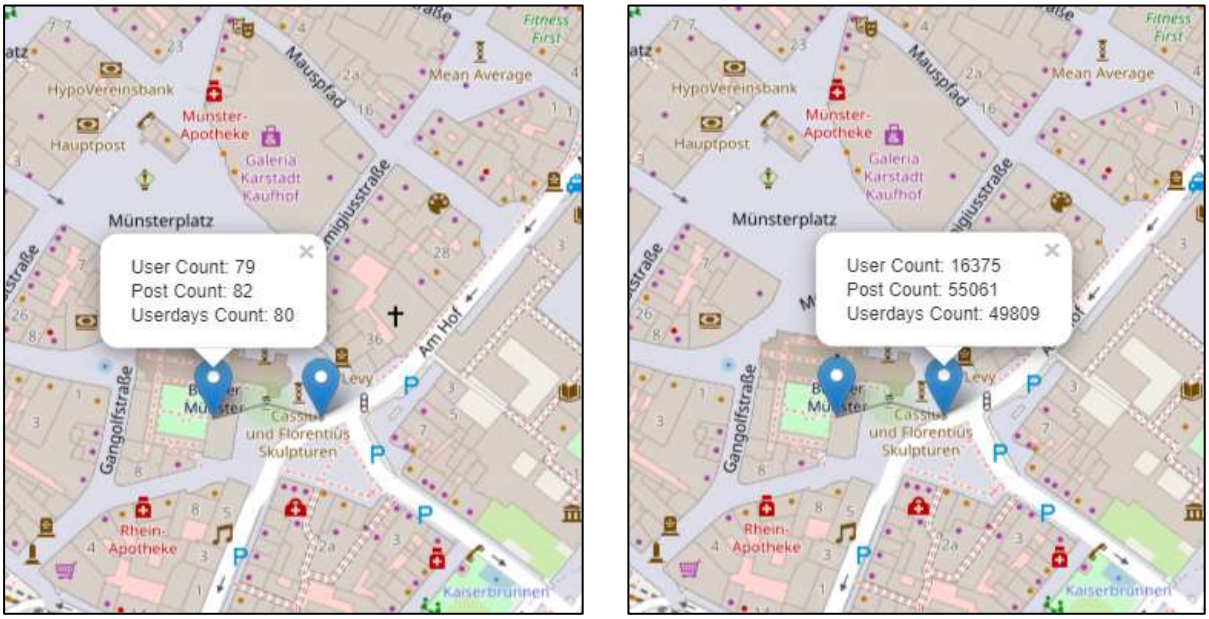


Figure 11: Metrics for the Instagram location "Bonner Münster" (left), aggregated metrics for all Instagram locations with <20m radius of "Bonner Münster".

[13] The real name cannot be revealed to legal reasons.

When aggregating location IDs to their coordinate as done in fig. 11, the enormous difference in posts (>55 000 vs. <100) is not visible at first glance but requires particular attention as otherwise it would lead to extremely biased results. This ambiguity – which can be even worse considering location IDs for city districts, regions, countries, continents etc. – is hard to overcome but most times, can be identified through their high variance in posts (empirically determined here for the data of Bonn). Usually, a rule of thumb is that the bigger the geographical area referred to, the higher the number of posts.

A Facebook location is mostly related to a geographical area in reality such as a park, a house or similar but very seldomly to a point which, in turn, makes it quite hard even for Facebook itself, to automatically understand what exactly a location coordinate refers to. A research team of Facebook acknowledged for example the "exceptionally challenging problem" of deduplication (DALVI et al. 2014: 409) and found that "location coordinates themselves are extremely noisy" (ibid.: 411).

Still, mining such locations and their posts at scale offers a high insight potential not only to city planners and citizens but also for commercial interests – even tough, particularly when commercial interests come into play – legal and ethical perspectives must be discussed beforehand.

As of time of writing, there is no direct research API access to either Facebook or Instagram. Still, as for the public nature of all Instagram posts (unless set differently) and locations, they can easily be searched manually[14] and mined automatically (see WECKMÜLLER 2021d) via their public API with certain limits.

3.3.1.2 Twitter

Twitter is a microblogging service and social network. One can post so-called tweets, containing images, text and tags limited to a maximum of 140 characters (MORSTATTER et al. 2013: 400). Lastly, TWITTER (2021: 2) reported 192 million monetizable daily active users (DAU).

Geotagging in TWITTER is either based on POIs of the company foursquare or can contain a custom location label and the exact GPS-Coordinate, but only when using Twitter's mobile app (TWITTER 2021b). While working with POIs, even though the POI creating rights are exclusively granted to a company, it technically leads to same issues as mentioned above for Instagram locations. However, for the posts, being tagged with a GPS-Coordinate, the precision in comparison to Instagram is higher and adds further detail and hence analysis opportunities.

In comparison to Instagram, few posts on Twitter, also called "Tweets", are geotagged. In fact, some studies talk of around 1% (MORSTATTER et al. 2013: 407), others come to a conclusion of around 2% excluding a varying percentage which can be approximately identified through time zones and text matching (BURTON et al. 2012: 1) or 2.3% in a large-scale study analyzing 40 billion tweets (HUANG

[14] Either the Facebook or the Instagram pages eventually lead to the same locations (see FACEBOOK 2021b & INSTAGRAM 2021b).

& CARLEY 2019: 367). While the exact numbers are irrelevant to this thesis, an approximate benchmark is sufficient to understand the value of Twitter data in the LBSM context.

Twitter offers limited access to its API for researchers after an application process (TWITTER 2021c). Such an account was used for this thesis. However, as generally Tweets are publicly available (BURTON et al. 2012: 2), just like Instagram posts, they could just as well be mined without direct access to the elsewise limited researcher's API.

3.3.1.3 Flickr

Flickr is an "online photo management and sharing application" where members can connect and interact (FLICKR 2021a: n.p.). It hosts more than 10 billion photos (ibid. 2015: n.p., as of 2015) partially with metadata such as timestamp and coordinates and counts around 60 million visitors every month (ibid. 2021b: n.p.).

In SM research it is well known for its open API which is freely accessible for everyone for non-commercial use (ibid.).

As of 2009 roughly 3.3% of all Flickr photos were geotagged[15] (FLICKR 2009: n.p.). Unlike Facebook or Instagram, Flickr provides a limited feature to search photos via a map interface (see FLICKR 2021c: n.p.).

3.3.2 Demographics

As mentioned earlier, the social facet can partially be represented through the LBSN being used.

BURTON et al. (2012: 2f) note that the user groups of the individual LBSNs are different in nature, i.e. they can differ according to classic demographic characteristics such as age, gender and ethnicity.

In their yearly telephone survey with 1502 U.S. participants, the PEW RESEARCH CENTER (2021: 7) found that SM usage in 2021 "varies – sometimes widely – by demographic group". "Adults under 30 stand out for their use of Instagram, Snapchat and TikTok" whereas Facebook is used equally among different age cohorts below 65 (ibid.: 7). Depending on the LBSN, the user cohorts thus differ greatly (BOYD & CRAWFORD 2012: 669).

As SM users tend to be younger (PEW RESEARCH CENTER 2021: 7), LBSM data consequently do not represent the entire population. Depending on the context, this can be seen either as a weakness, if a holistic picture of reality is to be created where possible, or as a strength, in that the needs of the relevant cohorts can be specifically identified (BURTON et al. 2012: 3).

[15] There were no newer statistics available at the time of writing.

Building on this insight, a post that comes from a certain LBSN can therefore be assigned to a cohort with a certain probability or put differently, a set of posts represents a certain cohort to a certain extent. However, user demographics are not discussed further here.

3.3.3 Legal and Ethical Discussion

3.3.3.1 Legal

There are different aspects where legal challenges must be considered. Starting from data mining and data processing as well as saving to publishing require individual legal discussion.

The process of data mining describes the act how the data get from the data collector i.e. LBSN to the data miner i.e. researcher (fig. 10). As mentioned in the previous chapter, there are three main variants:

1) API
2) Manual search and analysis
3) Automatic data scraping

If an API is used, this can usually only be done with an authorized key and is therefore completely legal. This is the case, for example, for Twitter and Flickr in this thesis.

For Instagram – in the absence of an API for researchers – the public API was used, which otherwise delivers the public data to the browser (see WECKMÜLLER 2021d). Since no technical hurdles are overcome or keys are hacked, this can be described as legal. There are no court rulings or similar on this matter in the non-commercial sector.

Manual viewing of public posts, i.e. regular use of LBSN, is necessarily legal as this is how LBSN work.

Data scraping refers to the machine reading of public data which is not used in this thesis. The legal situation is more complicated but is not discussed here due to lack of necessity.

The storage and use of personal data in contexts other than the intended one usually require the consent of the concerned user. Furthermore, Instagram explicitly prohibits, for example, applications that are based on its content and exploit it in another context: "You can't modify, translate, create derivative works of, or reverse engineer our products or their components" (INSTAGRAM 2021a, n.p.).

Strictly speaking, then, applications such as those by LIN et al. (see 2021; 2016) violate the terms of use as they use Instagram data in their public web app.

However, since at no point the original data are stored in the finished dashboard[16], but only a probabilistic abstraction of it and, moreover, broken down into its atomic components, the data no longer fall

[16] The step from retrieving an API to the finished HLL set is done completely in-memory. For details see ch 3.3.4.

under the corresponding strict classification of the GDPR of raw data but under statistical data (DUNKEL 2020: 7; fig. 8). In this way, the dashboard does not violate the Instagram terms of use.

For statistical data, for example, article 17 of the GDPR (2016: art. 17), the "Right to be forgotten" does not apply. For this reason, they can also be fully processed, i.e. for the purpose of storage and ultimately also publication.

REYMAN (2013: 524) argues that legally there is still a dividing line to be drawn between "user content and user data". User content (especially visual material such as photos and videos) is still subject to copyright laws, whereas user data, i.e. meta data in particular, are not additionally protected. The question to what extent meta data, e.g. picture captions, hashtags etc. are protected by copyright laws is still open and there is no legal precedent.

Disclaimer: This chapter is to be understood merely as a technical classification and therefore does not guarantee legal certainty.

3.3.3.2 Ethical

The ethical questions cannot be answered definitively neither. A general discussion is needed in society about the extent to which data may be used for a purpose other than that for which it was intended. This subsection is intended to provide a starting point in this regard.

The following discussion refers to a dashboard that implements all measures presented in ch. 3.2.8.

A quotation is prepended to the discussion:

> "Because of the public nature of the tweets, users do not have any expectation of privacy, so researchers may openly observe the content" (BURTON et al. 2012: 2).

This is an extreme and one-sided view, in terms of the public nature of tweets, which is to be explicitly disagreed with here. Unfortunately, according to METCALF & CRAWFORD (2016: 2) this is a common argument in research. Rather one should consider the following: "Just because content is publicly accessible does not mean that it was meant to be consumed by just anyone" (BOYD & CRAWFORD 2012: 672). DI MININ et al. (2021: 437) even demand that

> "[…] similar to any research involving people, scientific investigations based on social media data require compliance with highest standards of data privacy and data protection, even when data are publicly available".

They emphasize that "the risks to individual users' privacy and well-being can be substantial "as once leaked, such data have "the potential to cause psychological or physical harm to an identified person" (ibid.).

Users often upload their data to LBSN without considering possible consequences (KESSLER & MCKENZIE 2017: 7) or without being aware of being the subject of research (METCALF & CRAWFORD 2016: 10). As described in ch. 2.4.2, so-called inferences (BAROCAS & NISSENBAUM 2014: 55) can be used to combine different data sets in order to gain information that in its sum leads to new insights and thus might significantly harm the user's privacy (METCALF & CRAWFORD 2016: 2). If a user has uploaded a post in the belief that the personal data would not be used for anything other than the corresponding LBSN, it might pose a risk if it is suddenly used against the user, e.g. to check attendance at work. For this reason, the balancing act between a thoroughly existing research interest and the privacy of the user is not easy and can by no means be answered unequivocally.

In general, ethics and privacy are often forgotten in research. A negative example on a considerable scale is a study by HOCHMAN & MANOVICH (2013), for example, who republished millions of users' pictures in high resolution without their consent and without self-critical reflection. Unfortunately, such studies are not exceptions, but are seen as a general trend or "growing discontinuities between the research practices of data science and established tools of research ethics regulation" (METCALF & CRAWFORD 2016: 1).

Basically, for this dashboard, put simply, the interest in the absolute privacy of the individual and personal data sovereignty must be balanced with the research interest or, in this case, the societal interest in improved information to reduce spatial inequality through improved resource allocation.

Although there is a certain, albeit minimal, residual risk of an attacker obtaining the salt key, the hash function and access to the database even when using HLL (ch. 3.2.8), it cannot be completely ruled out that a potential attacker could derive information from the data and thus violate the privacy of individual users.

However, it should be pointed out once again that an HLL set would be completely worthless to a potential attacker without a salt key and hash function or vice versa. In addition, only small sets about the atomic components of posts are available and the data are public, which makes an attack (except in rare cases when a user has deleted a post afterwards) usually pointless.

Still, the question whether data can be used for different purposes than originally intended is highly controversial and needs a careful evaluation before every usage. As proposed e.g. by ILIEVA & MCPHEARSON (2018: 561), the Code of Ethics for GIS professionals (see GIS CERTIFCATION INSTITUTE 2003) provides a useful "starting point" for such questions.

Another important question is to what extent data are interpreted correctly in the end. As BOYD & CRAWFORD (2012: 666f) put it, "claims to objectivity and accuracy are misleading" as "working with big data is still subjective" and they still need to be interpreted. If this fact is ignored, the "biases and

limitations" are not understood, and there is a complete reliance on a falsely assumed "objectivity" (ibid. 667) of the data, "misinterpretation is the result" (ibid. 668).

Like this, inequality could be reproduced, very contrary to the original intentions of a LBSN dashboard for MSGG as, in the worst case, "potential harms of data science research are unpredictable" (METCALF & CRAWFORD 2016: 1). Put in a nutshell, this can be called "open data gone wrong" (ibid.: 9).

This thesis does not presume to provide a conclusive answer to these questions, but on the contrary would like to encourage readers to take a closer look at the following specific questions in particular, as the necessarily limited scope of this chapter cannot reflect the complexity of the topic appropriately.

1) Is the use of public but not for other purposes originally intended LBSM data justified in view of a possible societal informal value about different user cohorts to increase the quality of life and establish spatial justice?
2) To what extent is the risk of privacy violation acceptable?
3) How high is the societal value and does it justify the potential risk of privacy violation?
4) What are the risks of misinterpreting the dashboard data?
5) Is it possible to misuse the dashboard for criminal purposes? How likely would this be and would it be an acceptable risk?

3.3.4 Dashboard Setup

As there are supplementary GitHub repositories available with a concrete setup description (WECKMÜLLER 2021a; 2021b), the prototype is not explained in technical detail in this chapter. However, an idea of how the data are processed and delivered to the end user, clarifies all the previous chapters and is hence briefly described.

The first step is data retrieval from different LBSN. All these data are read into the HLL-DB with the help of lbsntransform (see DUNKEL & LÖCHNER 2021b) that can be regarded as an adapter for parsing different LBSN and data formats, i.e. comma-separated values (CSV) or JavaScript Object Notation (JSON) to the LBSN data scheme. With the help of lbsntransform and its SQL commands, the data are immediately split up into its atomic components and read into HLL sets into the HLL-DB.

This process can be integrated well in regular data mining pipelines, i.e. when the data are retrieved (e.g. via the Twitter or Flickr API) it is processed immediately in-memory with lbsntransform (see DUNKEL & LÖCHNER 2021c) so that the original data do not need to be saved.

When datamining, white- or blacklists can be used to avoid critical terms (e.g. offensive language) or just focus on a particular topic such as UGS by only creating HLL sets for particular terms, e.g. "tree", "park", "sun" etc.. This is highly recommended as it reduces not only the data amount but also the potential risk of privacy issues.

When reading data into the HLL-DB, the previously mentioned salt-key can be applied if suited (HLL unions would be prevented which are crucial to the following case study), a hash function and a seed is chosen. At this point, the HLL-DB contains certain, limited HLL sets and can easily be queried with SQL commands in psql or pgAdmin.

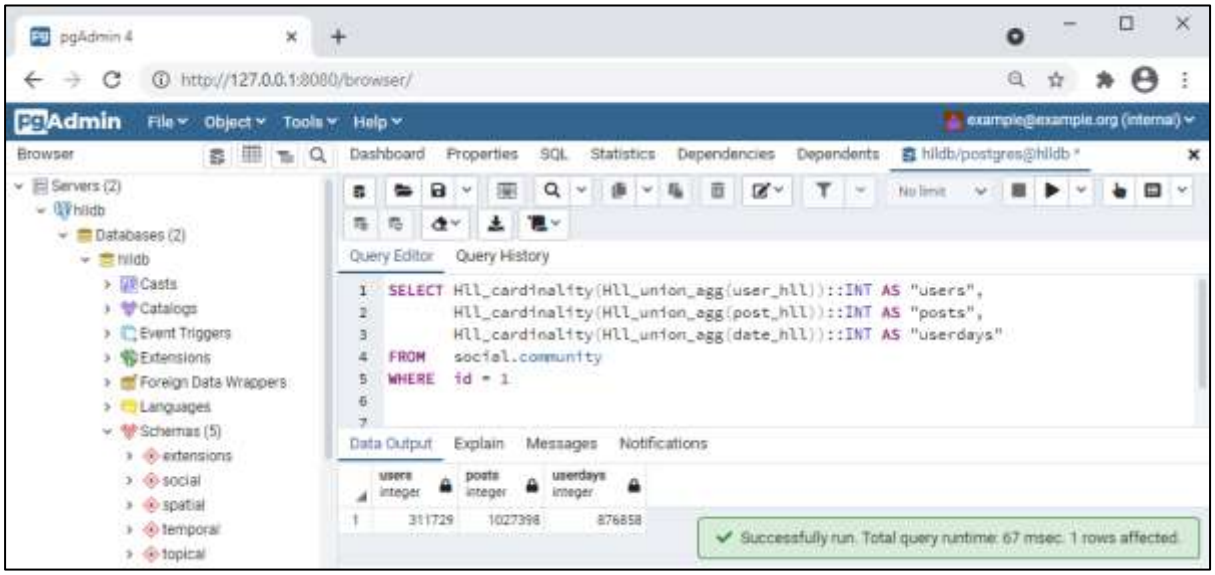


Figure 12: pgAdmin 4 screenshot with a sample SQL command for querying the overall number of distinct users, posts and userdays for Instagram in the entire database.

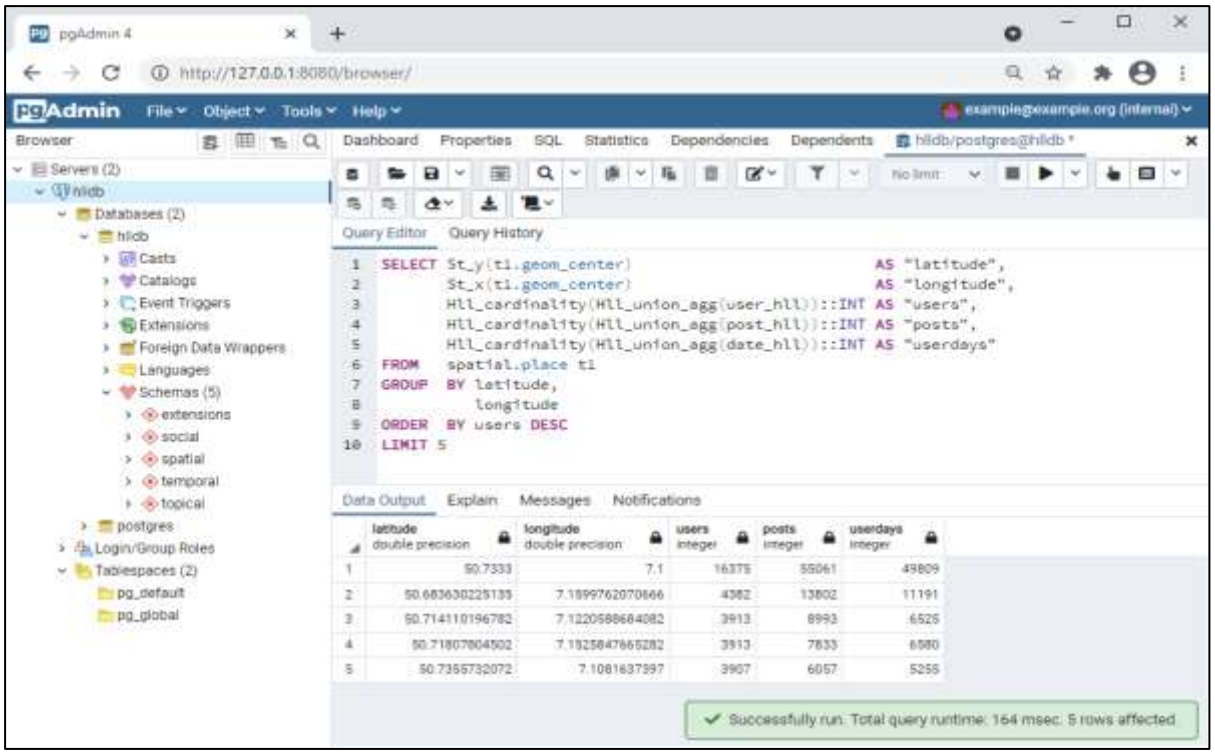


Figure 13: pgAdmin 4 screenshot with a sample SQL command for querying the top five metrics for all posts from different LBSN, aggregated for coordinates.

For development purposes the HLL-DB is yet only locally hosted (fig. 12 & 13; "127.0.0.1" is localhost) so that no one can access it. Fig. 13 already gives an impression of what data are sent from the back- to the frontend.

The backend is completed with a web framework of choice, i.e. fastAPI (see RAMÍREZ, S. & CONTRIBUTORS 2021) which handles minimum HLL set sizes (HLL sets must contain at least n elements), possible access control and further security measures.

The frontend is used to display custom queries on a map. For this purpose, only open-source packages were used, i.e. Leaflet (see AGAFONKIN & CONTRIBUTORS 2021) and various plugins.

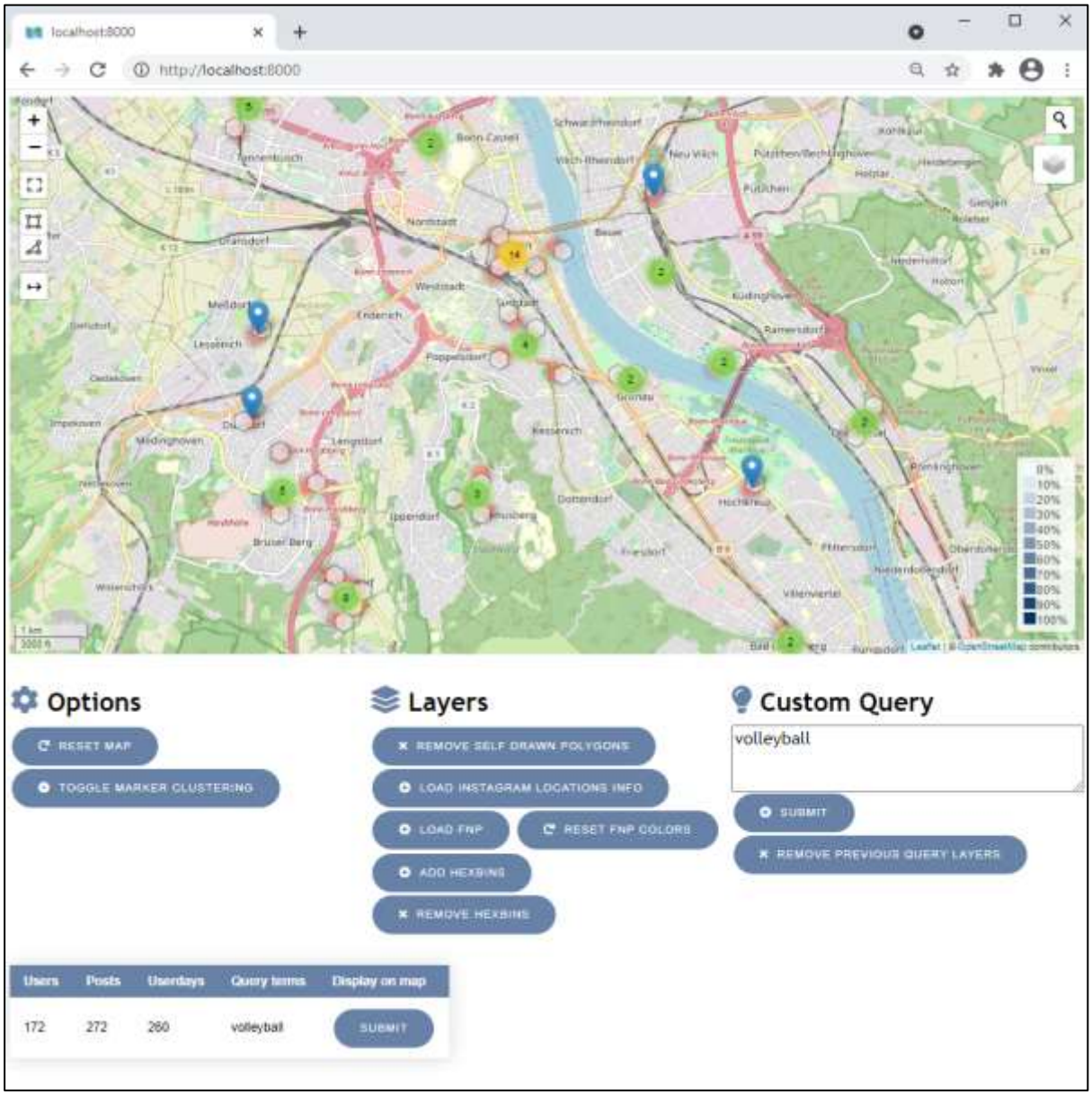


Figure 14: Prototype frontend with custom query results for the term "volleyball" in Bonn. Note: Pins represent the locations and are clustered on lower zoom levels (green circles). On top of a heatmap (orange colors) hexagonal bins (hexbins) represent aggregated counts.

Adding custom cascading style sheets (CSS), the prototype frontend results in an easy-to-use graphical user interface (GUI). Instructions on how to use it can be found in the respective documentary section on GitHub (see WECKMÜLLER 2021a).

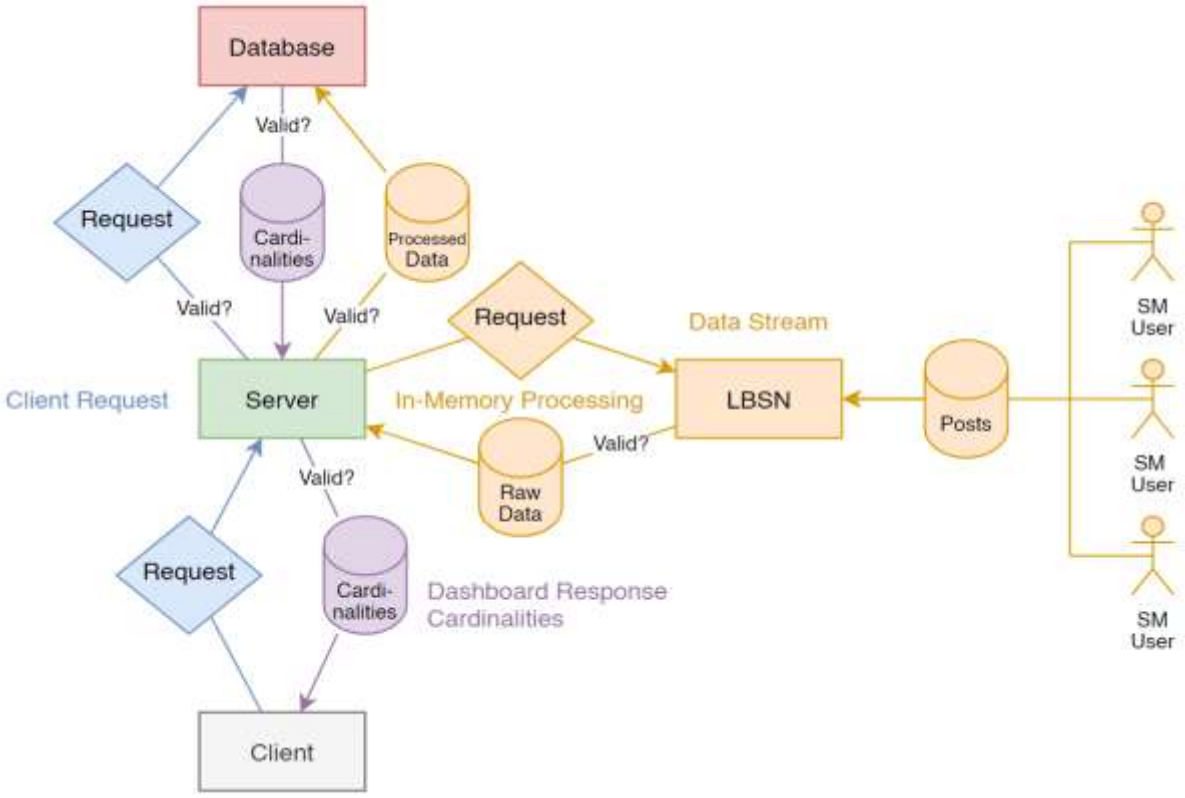


Figure 15: Data flow for a privacy-aware client-server architecture.

Fig. 15 sums up the data flow. While the yellow arrows indicate, how the LBSM posts are processed in-memory ("streamed") and processed for the database, e.g. daily, the blue arrows indicate a data request from a client, sending an initial request to the server for validation. The violet arrows indicate the data backstream if the request is valid, e.g. a minimum number of n posts is reached. The database sends the cardinalities to the server which in turn also validates them once again. If valid, it eventually sends them back to the client where the response cardinalities can be displayed in the frontend.

3.3.5 Functions

There is wide range of functions already implemented in the dashboard. The whole data flow (fig. 15) is implemented for the previously described LBSM platforms (ch. 3.3.1).

For the sole purpose of this thesis, the development of ready-to-use functions for every SM facet in every thinkable way was indisputably out of scope. However, the spatial and the thematic facet are fully implemented whereas the temporal and the social facet could easily be compensated by keeping more HLL sets e.g. for certain timespans or differentiated by LBSN. The status quo implementation fires the requests for the entire HLL-DB and hence for all LBSN. In this manner an additional layer of privacy

is introduced as the client cannot know, what share originates from each individual LBSN.

Figure 16: Spatial queries in the dashboard frontend.
From left to right: polygon, multipolygon, multipolygon with holes.

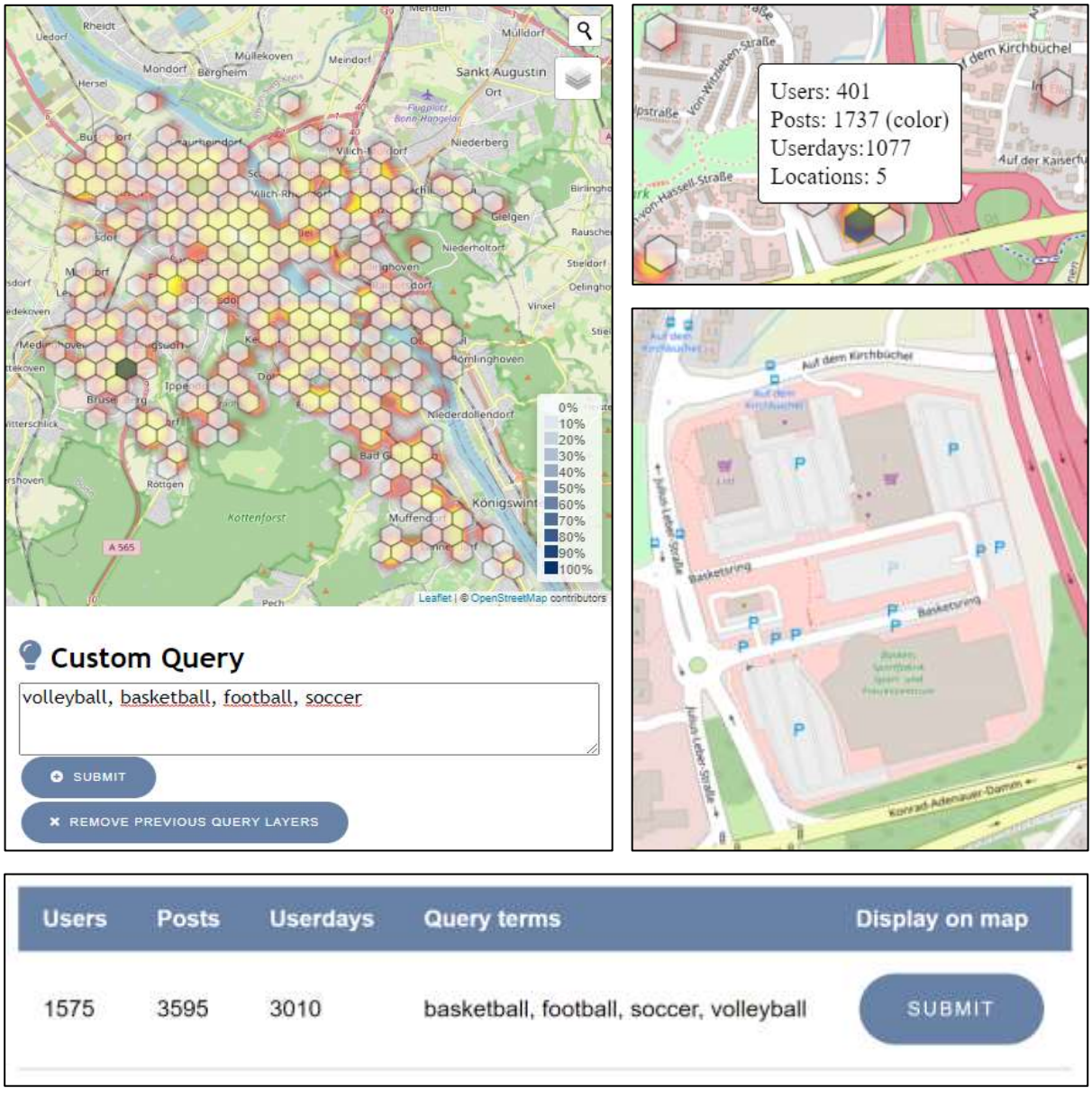


Figure 17: Screenshot compilation for thematic query for various sports in Bonn. From top left to bottom right: query results, zoom on hexagonal bin with most posts, discovering a big sports center, query metrics.

The two main functionalities which can be combined, are a spatial query and a thematic query.The spatial query can be performed by drawing a polygon, a multi-polygon or even a multi-polygon with holes (fig. 16).

The thematic query can be performed by searching for certain keywords (terms) that occur in the posts caption, like e.g. sports (fig. 17). In this way, one can quickly get an impression where to find hotspots of all kind. The request should happen at least in English[17] and German, since SM generally tend to multilingualism with a – in the Bonn context dominant – local language (German), English and other minority languages and a high affinity for "'intra-linguistic' variety of register and styles" as shown in a case study by LEPPÄNEN & KYTÖLÄ (2017: 161).

3.3.6 Plugins

The front- and backend are built in a way that it is easy to add plugins.

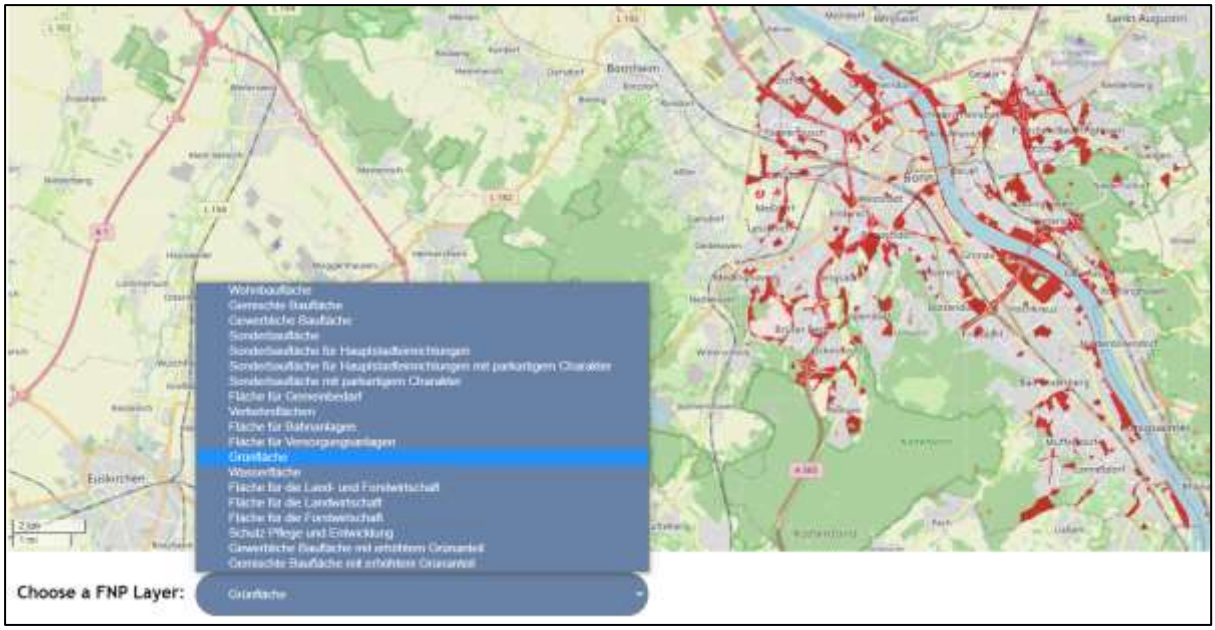


Figure 18: Dashboard plugin for choosing land use category of Bonn.

As can be seen in fig. 18 such a plugin is already provided. The button "LOAD FNP" (fig. 14) loads the land-use plan (DE: "Flächennutzungsplan" or "FNP") and provides further analysis possibilities. In fig. 18, a category such as "Green Spaces" (DE: "Grünfläche") can be selected. The geometries are then displayed and its metrics analyzed just like in fig. 17. In the same way, any other geometries could be queried which opens a broad range of application possibilities. For example, public participation could

[17] For a comparative socio-linguistic comparison and the role of English in international SM see DAILEY-O'CAIN (2017).

be encouraged via the same interface, e.g. by allowing public comments on locations or including already existing citizen petitions.

As mentioned previously, an interface for adjusting the temporal and social facet is yet to be implemented.

4 Case Study Bonn

This chapter demonstrates for which purposes the dashboard can be used and what potential it holds. It serves as an application-oriented guideline as to which phases of the dashboard can be implemented in a sequential manner by treating individual use cases that increase in complexity with a partial conclusion (ch. 4.5) and extension options in (ch. 4.6).

Phase 1 (ch. 4.3) includes the simplest queries for which HLL sets do not need to be intersected, i.e. either simple queries that refer to a facet or aggregates that are formed in lbsntransform and the HLL-DB by default. In this phase, by reduction to simple queries, the individual facets are presented through a very open research question: **Which patterns can be identified on LBSM for the whole city of Bonn?**

In phase 2 (ch. 4.4), specific queries become possible through certain base combinations (ch. 3.2.7) and HLL intersections. As a result from an earlier cooperation and the concrete interest of the city of Bonn, Bonn's UGS are analyzed.

The aim is to show the potential and give an idea for whom such a dashboard would be useful whereas the outlook is to be understood as a future phase 3 (ch. 4.6) and provides ideas on how the dashboard can be extended through possible plugins.

4.1 Study Area

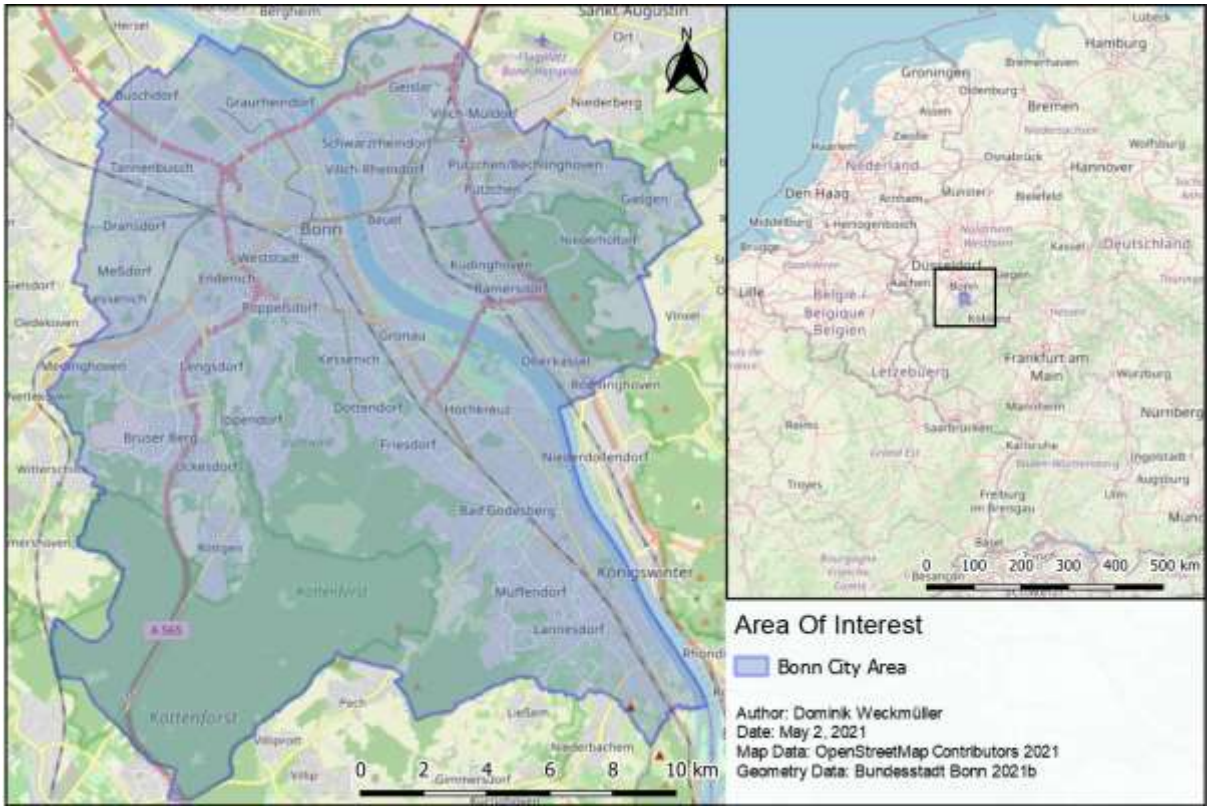


Figure 19: Area of interest for the case study (official Bonn city area geometry, see map for sources).

For various reasons, the geographical focus of this work lies on the administrative area of the German city of Bonn (fig. 19).

An important reason is the "spatial local knowledge" of the author, which shall be defined as "spatially referenced personal knowledge" (RANTANEN & KAHILA 2009: 1983). Since "local knowledge is very versatile" (ibid.: 1989), it was used to validate results and to question initial contradictions in the data. This particular knowledge is exceedingly useful, especially in an initial data review, and helps in interpreting the results.

Since this thesis is very application-oriented and the goal is a functioning prototype that wants to be tested in practice, another important reason is the cooperation with possible administrative partners, i.e. municipalities. Representatives of the city of Bonn have signaled their interest in advance, therefore the study area was specifically selected accordingly. In addition, the geometry data of the Bonn land use plan was provided in order to functionally integrate it into the dashboard.

In addition, another goal was to incorporate data from different platforms, which requires that posts exist for different LBSNs and the study area. This was ensured for Bonn, as a city with more than

330 000 (BUNDESSTADT BONN STATISTIKSTELLE 2021: 4) inhabitants and ca. 945 000 yearly tourists (BUNDESSTADT BONN 2021c: n.p., as of 2019).

4.2 Data

Because not every SN can be equally called LBSN, a certain spatial potential had to be present, which was checked for the most common SNs. Subsequently, it was examined whether and, if so, how easy it was to mine data from the respective LBSNs.

Here, the LBSNs Instagram, Flickr, and Twitter (ch. 3.3.1) were found to be suitable. These three LBSNs were also chosen in comparable studies for the same reasons, leading to a certain degree of comparability between case studies (cf. TENKANEN et al. 2017). The posts were mined via the corresponding APIs.

In order not to compromise the privacy of the users due to some high-resolution analyses in this chapter, an attempt was made not only to include posts from different LBSNs, but also to cover different time periods as far as possible. Apart from a rough period of approximately eleven years between 2010 and 2021 and the annual and monthly listing (ch. 4.3.2), it is not defined in more detail here from when the data per LBSN originates, since the temporal as well as the social components in this case study are largely excluded from the analysis. The bulk of the data comes from the LBSN Instagram for the reasons mentioned in ch. 3.3.1.1.

Table 11: Metrics of sample data by LBSN with an error of 3-5%.

LBSN	Users abs.	Posts abs.	Userdays abs.	Users rel. [%]	Posts rel. [%]	Userdays rel. [%]	Posts per User[18]
Instagram	142 923	478 862	454 079	92.6	73.5	89.7	3.35
Flickr	4933	117642	22 929	3.2	18.0	4.5	23.85
Twitter	6450	55 478	29 215	4.2	8.5	5.8	8.6
Total[19]	154 306	651982	506 223				

Tab. 11 provides a rough overview of how many posts from each LBSN are included in the case study. It should be mentioned at this point that the absolute numbers in the LBSN comparison are not representative, i.e. the numbers do not allow the respective popularity to be quantified. The dashboard is

[18] The posts per user ratio differs greatly in the respective LBSN samples. However, this is not discussed further here.

[19] Since users on different LBSNs have different user IDs, these are also read into the HLL sets as different users. The unit is therefore unique users per LBSN. Therefore, no statement can be made here about whether there are any duplicate users. Consequently, the number of physical unique users tends to be lower in reality. The same applies for userdays and – as sometimes users tend to post the same content on multiple networks – also to posts.

completely platform-independent and can incorporate data from new platforms at any time through the package lbsntransform with the help of so-called mappings (see DUNKEL & LÖCHNER 2021d).

In general, with reference to BOYD & CRAWFORD (2012: 666f) and their warning about misinterpretation of data, it should be repeated and insistently emphasized that the entire population is by no means represented here, but only very specific population groups (ch. 3.3.2).

Due to limited data availability and the sensitivity of personal data, the social facet here cannot be broken down further than by LBSN and could be subject to further research.

Thus, the analyses and results are to be understood only as a very specific, non-representative reflection of what is reflected in LBSNs. Nevertheless, this work – in view of the constant danger of spatial injustice and under careful consideration of the risks as well as the ethical background and the potential benefits – is to be understood as a plea for the limited use of such data, which is now clarified by showing the analysis potential.

4.3 Phase 1 – Simple Queries for Bonn

4.3.1 Spatial

Spatial queries are elementary for an LBSN dashboard. While the spatial facet – just like all other facets – can be considered right at the beginning, i.e. when streaming the LBSN data, the focus here is on how spatial subqueries can be performed in the dashboard.

Since only a rough spatial limitation of the data is usually done during the datamining, as is the case here for the study area Bonn, smaller subsets of the data are of interest for various questions. This is illustrated here by the general spatial distribution of LBSM posts in the city of Bonn.

First, by drawing a rectangle, the metrics from fig. 14 & 17 can be displayed to obtain a general overview of the entire city area.

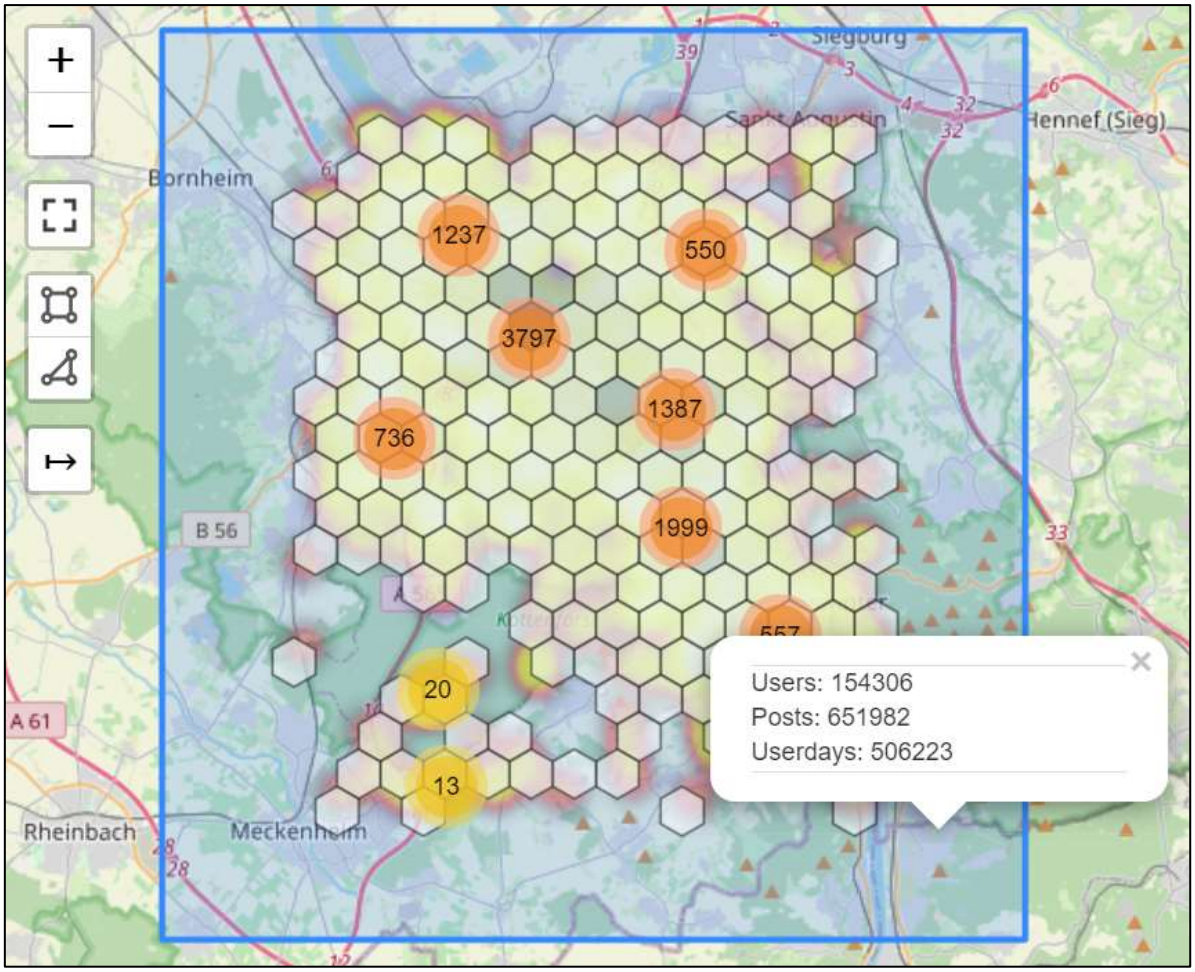


Figure 20: Case study metrics for Bonn.

As can be seen in fig. 20, the complete dataset totals to 651 982 posts from 154 306 different users on 506 223 different userdays with an error rate of around 3-5% depending on the settings due to the probabilistic character of HLL (DUNKEL, LÖCHNER & BURGHARDT 2020:7). Since these are only absolute numbers derived from the sample size, the numbers themselves do not tell anything about reality.

The higher the ratio of posts per user, the more the same users tend to post frequently. Conversely, a low ratio means that there are many "one-time posters" or people who only post a little, which could be an indication of tourists.

The metric of userdays indicates whether users tend to post on a few or many different days. In conjunction with the posts, this can be used to create a new metric. The posts per userdays, represents how many posts a user sends on average per day.

In this example the number is $\frac{651\,982\ posts}{506\,223\ userdays} = 1.29\,\frac{posts}{userdays}$. This number could be monitored over time or compared with other cities. The number shows that on average, when users decide to post

something, instead of just one, they will send 1.29 posts on that same day which requires further interpretation.

In addition to these standard metrics, which can be displayed for any area, the map automatically displays three additional layers.

Figure 21: Three dashboard layers: heatmap (left), locations (middle), hexbins (right).

A heatmap becomes lighter in the places where there are the most posts (fig. 21, left). The marker clusters unfold when zooming in or by switching manually (fig. 21, middle). The hexbins additionally show where the most has been posted, relatively speaking (fig. 21, right).

4.3.1.1 Practical challenges – Instagram locations

At this point, the specific problem of Instagram locations, which was treated theoretically beforehand (ch. 3.3.1.1), reveals itself in practice. These become problematic by their seemingly precise coordinates, which in reality refer to an area, which is larger than the desired analysis level.

Concretely, this means with respect to fig. 20 & 21, that the hexbins often do not reflect LBSM reality. For example, the darkest bin by locations does have the highest post count. However, this is only due to the fact that the coordinate for the most frequently used location on Instagram for the entire city of Bonn is located there[20].

For this purpose, Facebook has introduced a parameter to the locations that indicate whether there is an exact match for the location on a certain spatial level[21], but this is of no use if this information is not provided by the user who created it.

In practice, this means that appropriate care must be taken when interpreting the data. Even better is an automatic filter when streaming the data into the database. However, such a filter has not yet been

[20] Location ID: 107481562. API-Request: https://www.instagram.com/explore/locations/107481562/?__a=1 (last access: 03/05/2021).
[21] The exact match can refer to country, region or city. Found in node "address-json" under above link ("exact_city_match": true).

implemented here. Instead, the locations were filtered manually depending on the use case and the respective AOI.

4.3.1.2 Spatial distribution in Bonn

To understand the spatial distribution, different zoom levels can be selected. All layers automatically adapt to the corresponding level and become more detailed when zooming.

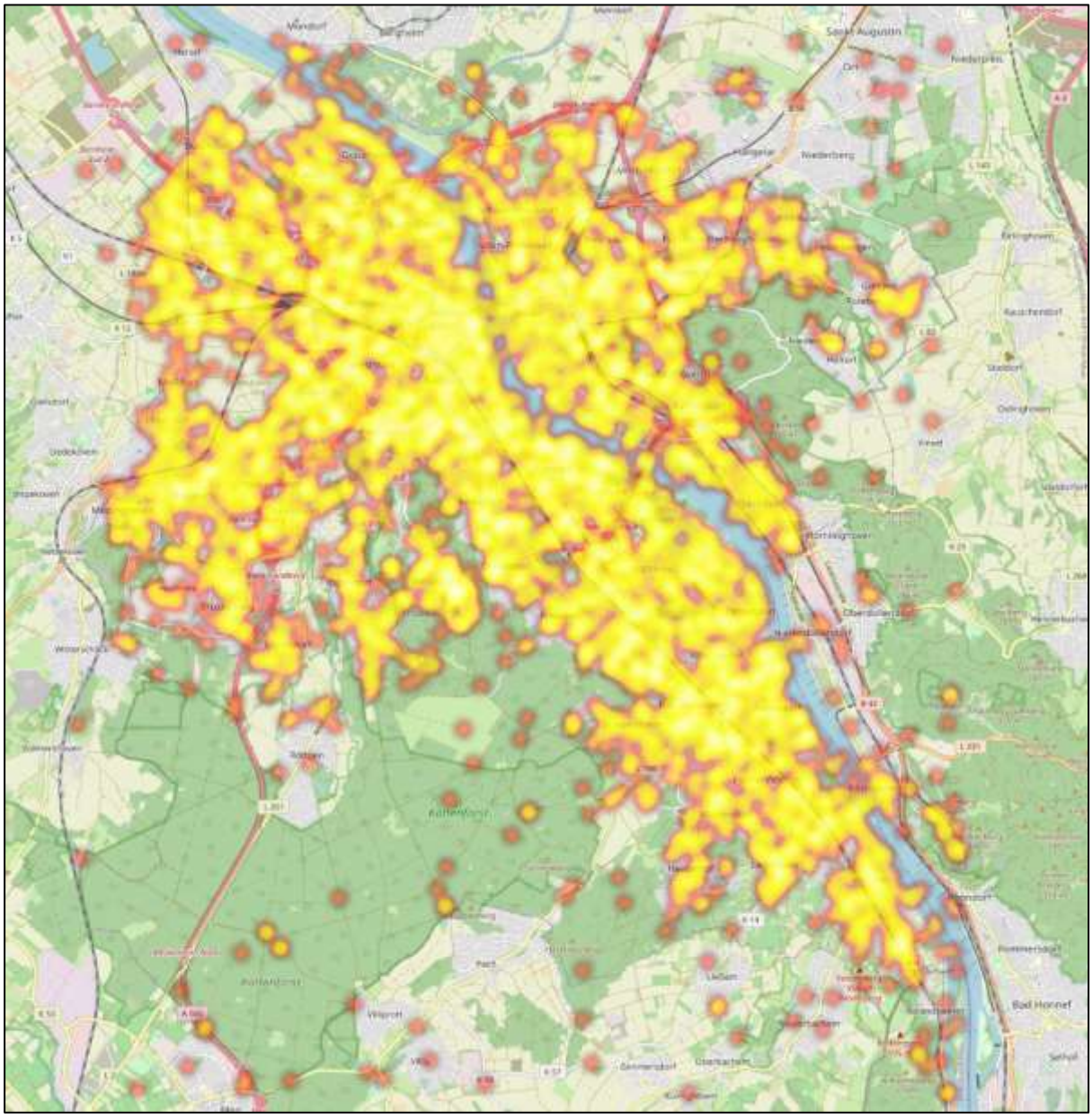


Figure 22: Heatmap for Bonn.

Fig. 22 is a first, simple heatmap impression of where most LBSN locations can be found. It is easy to see that very little is posted in the „Kottenforst" – a large forest area with recreational trails in the south-

west of Bonn – compared to the more urban areas in the center. At this zoom level however, clusters are hard to recognize.

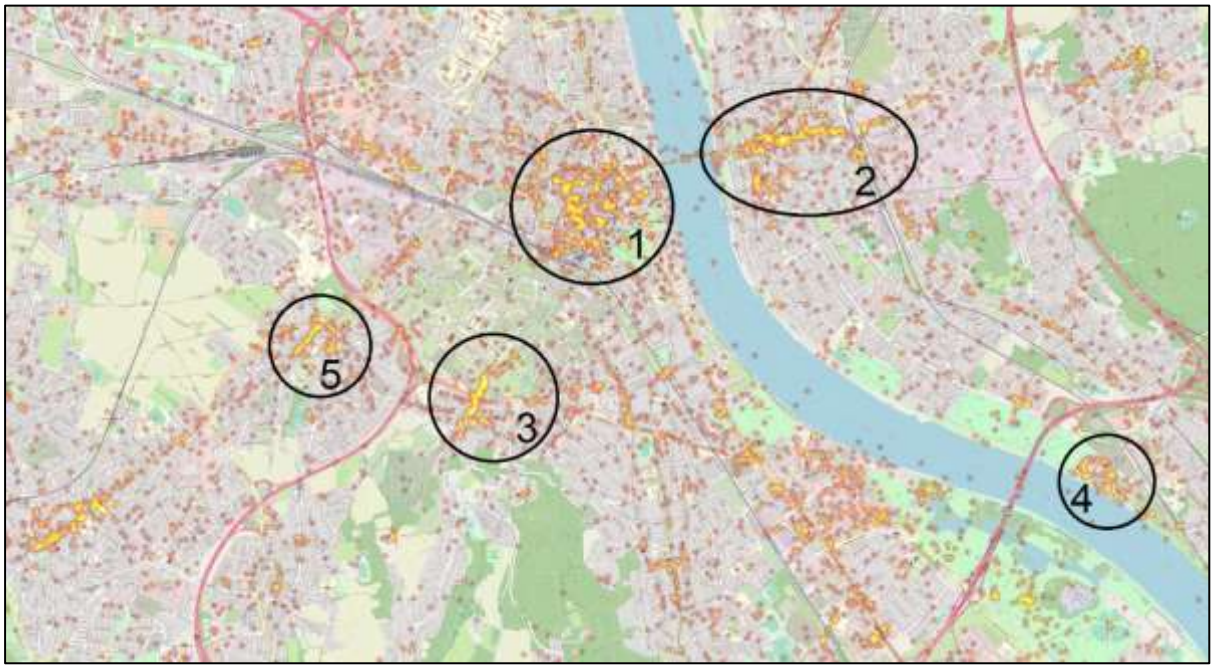


Figure 23: Heatmap on high zoom level with five sample clusters.

However, when zooming in further, one can clearly see different clusters (fig. 23). The color intensity of the heatmap is generated on the basis of the number of locations that have different coordinates, i.e. if there are many different locations but they have exactly the same coordinate, they are counted as one location and the metrics are aggregated to practically counteract the aforementioned confusion of Instagram's user-generated locations and their possible duplicates. The heatmap thus shows solely where certain agglomerations of locations are situated.

Table 12: Attraction factors for sample clusters.

Cluster no.	City district	Particularities and hotspots
1	Center	Shops, restaurants, daily market, attractions
2	Beuel	Restaurants, theaters, art centers
3	Poppelsdorf	University campus, Poppelsdorf Palace, botanical garden, restaurants, bars,
4	Beuel-Oberkassel	Business and research cluster, e.g. German Aerospace Center (DLR), Telekom
5	Endenich	Theaters, restaurants, bars

A brief, incomplete overview is given for the locations occurring in the example clusters marked in tab. 12 based on the author's local knowledge. Since the heatmap does not indicate how many posts have been sent at the locations and thus how important a location is on LBSM, an additional visualization method is needed.

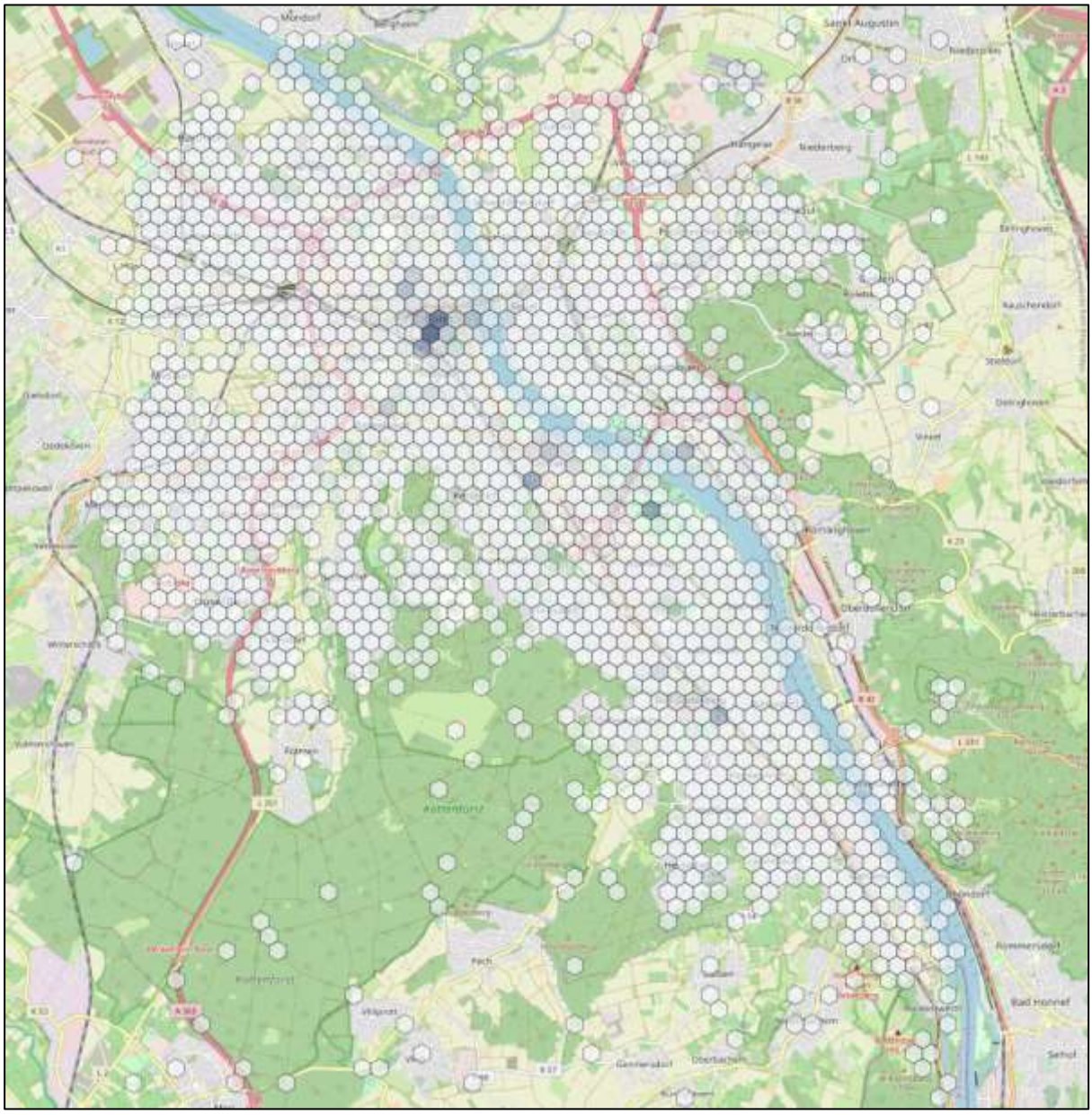


Figure 24: Hexbins for Bonn.

Aggregation bins are very suitable for this purpose. Hexbins were chosen here because they cover the area entirely and can be easily scaled for zooming (fig. 24). In addition, their radius and color intensity make them well suited to display two variables at once, which is explained later. On the technical side,

there is an already existing Leaflet plugin available which was modified for usage with HLL for this thesis (see WECKMÜLLER 2021e).

Each of these clusters from fig. 23 can now be examined more specifically with the help of the spatial selection, e.g. a part of Bonn-Poppelsdorf.

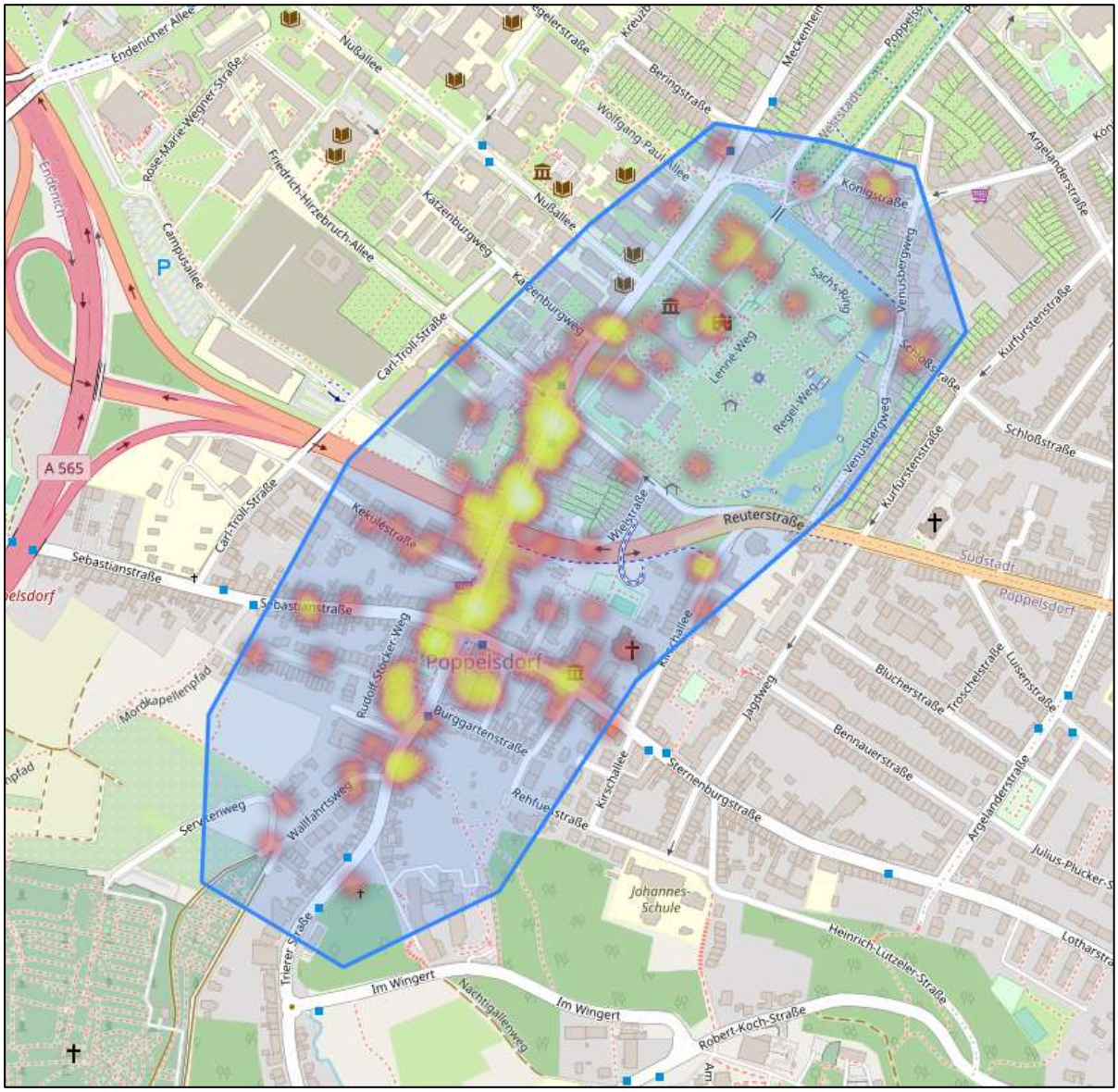


Figure 25: Custom sample AOI for Bonn-Poppelsdorf.

From a simple heatmap on high zoom level it becomes clear that many different locations are situated in the „Clemens-August-Straße" (middle of the cluster) indicating – put in other terms – a high aggregation density of certain POIs (fig. 25).

Table 13: Gastronomy OSM tags (OpenStreetMap Wiki 2021: n.p.).

Symbol	Description
	Restaurant (except fast food) / food court
	Cafe
	Fast food restaurant or snack bar or sandwich bar or similar
	Bar
	Pub
	Ice cream shop

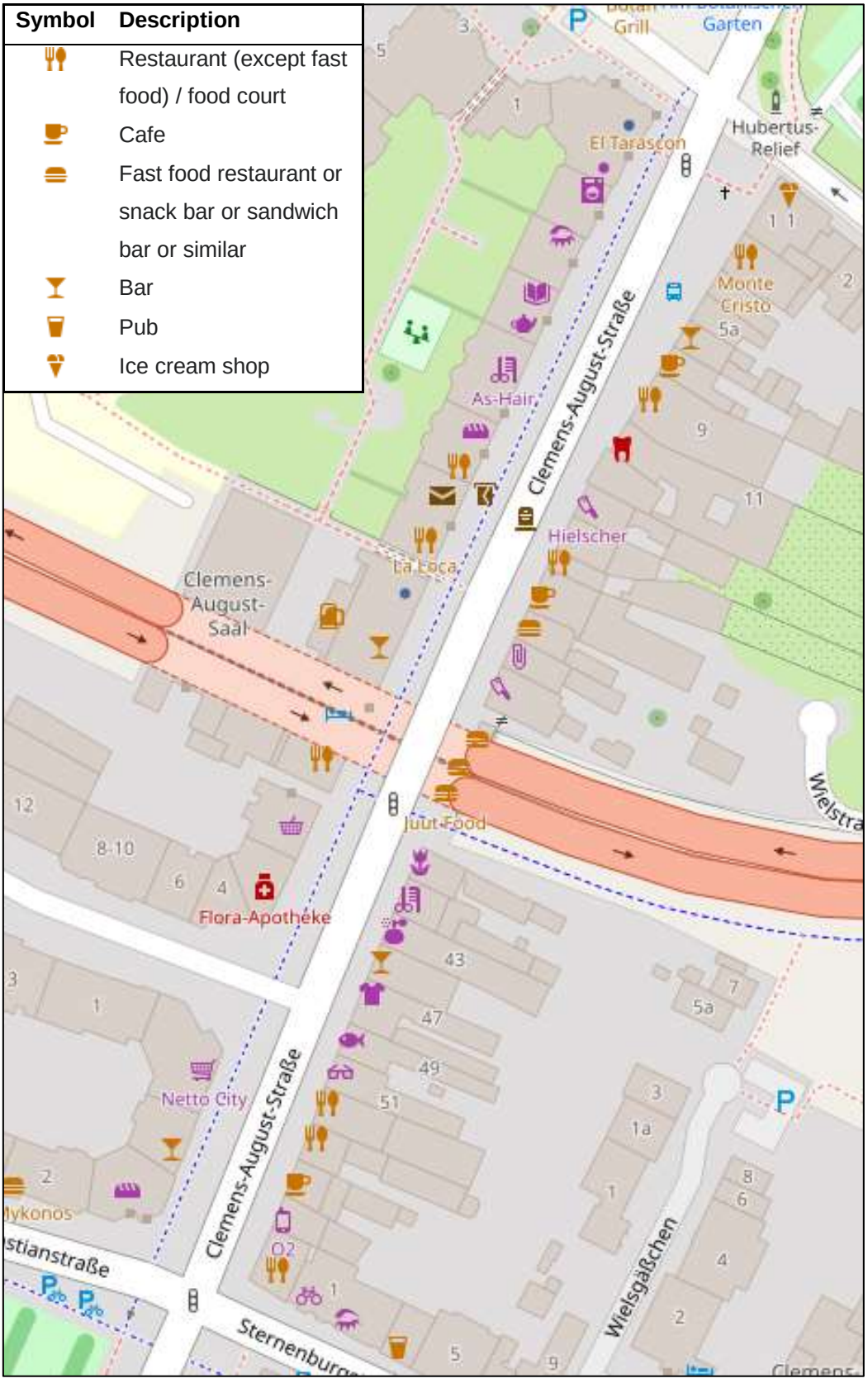


Figure 26: OSM tags for „Clemens-August-Straße“ Bonn.

In fact, a look at the highest zoom level in OSM reveals what is attractive about “Clemens-August-Straße”. For example, on the OSM background map (fig. 26) one can tell many restaurants and bars as indicated by the respective legend describing only gastronomy (tab. 13), but there are numerous stores in the street as well.

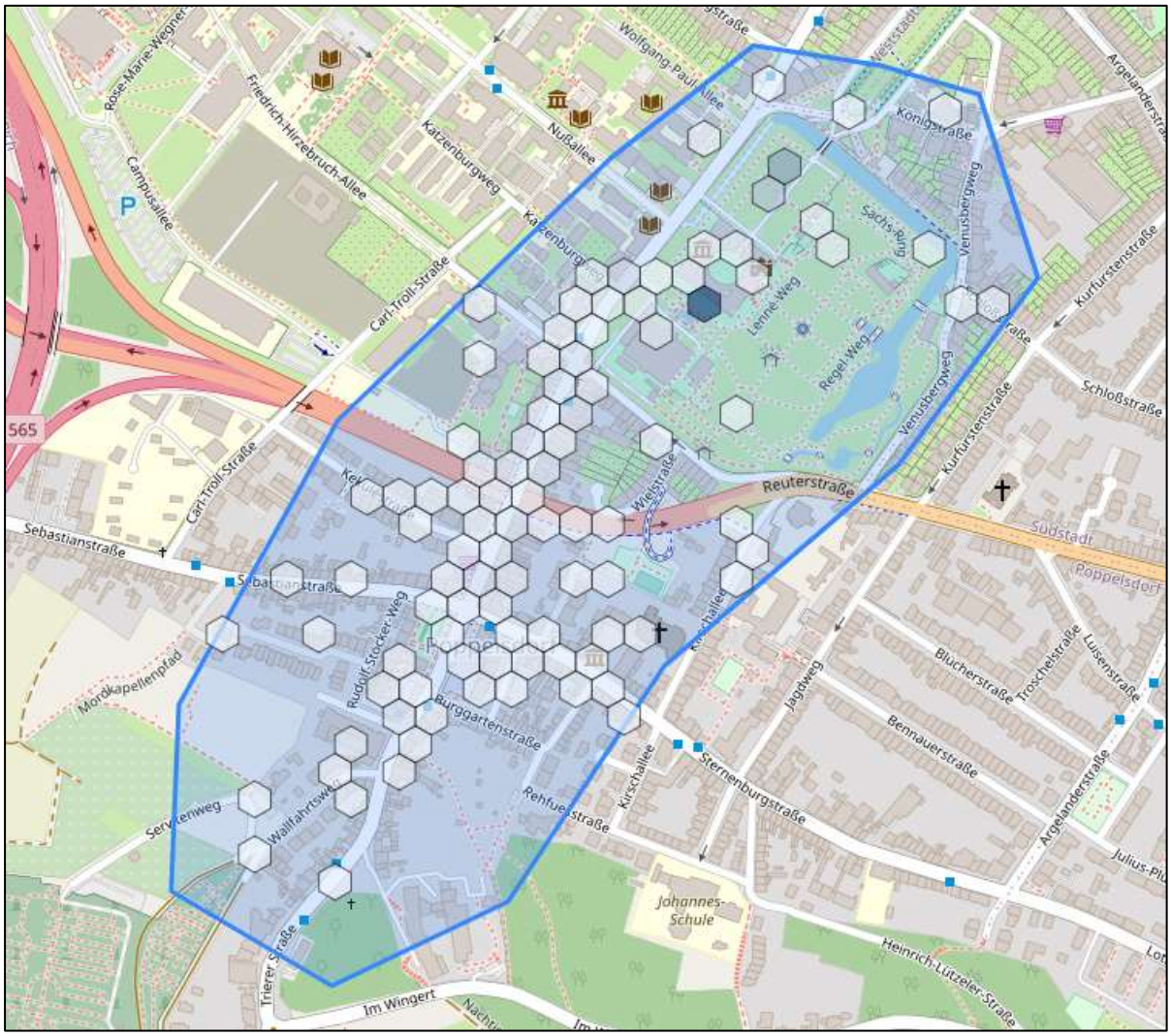


Figure 27: Hexbins for custom sample AOI for Bonn-Poppelsdorf.

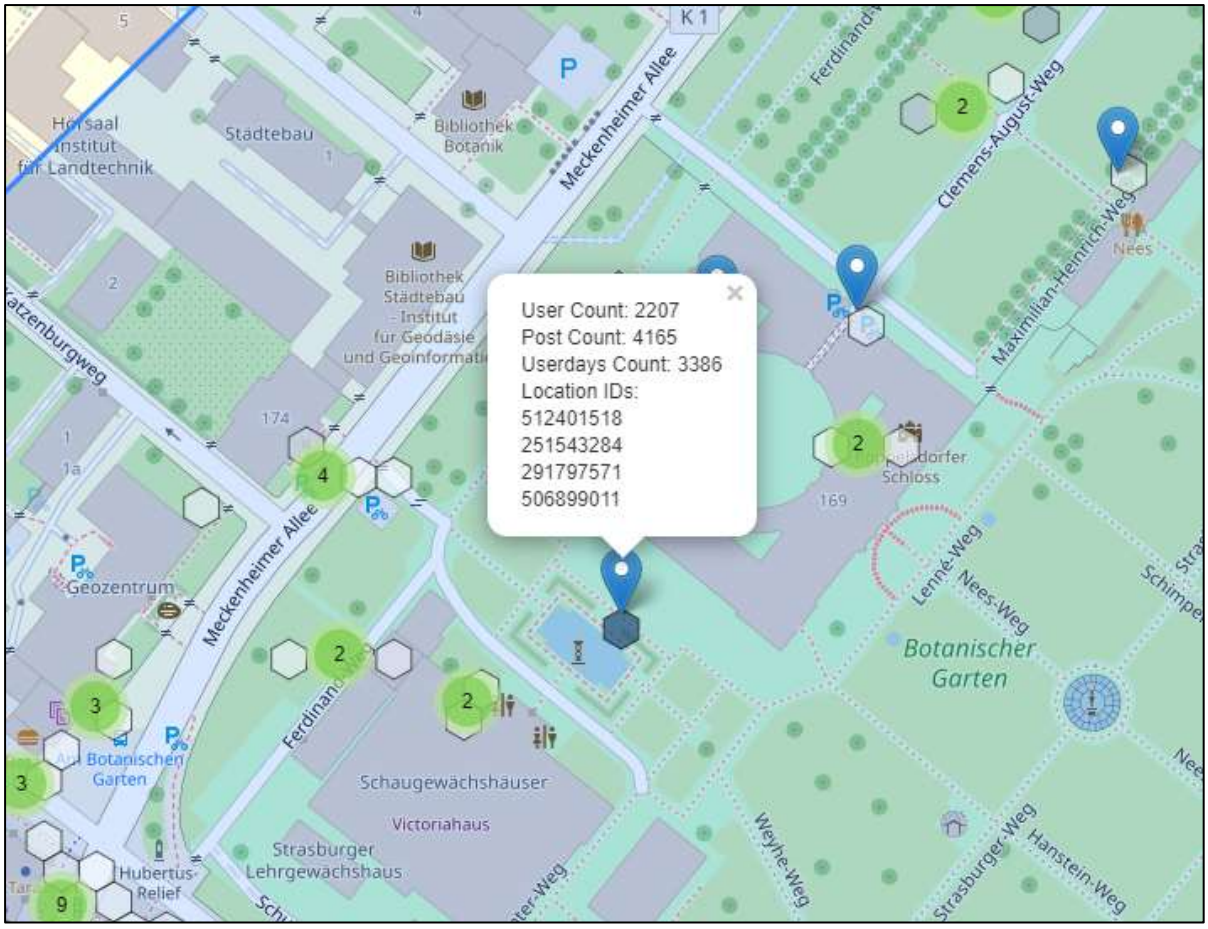


Figure 28: Metrics for custom sample AOI Bonn-Poppelsdorf.

In order to recognize where most posts originate from, hexbins are suitable. While there are many locations along the road (fig. 27), by far the most attention on LBSM is given to the botanical garden (fig. 28, dark blue hexbin).

Table 14: Metrics for botanical garden and custom sample AOI Bonn-Poppelsdorf.

	Users	Posts	Userdays
Botanical Garden (including 4 locations see fig. 28)	2207 (34.8%)	4165 (28.9%)	3386 (29%)
AOI Total (Bonn-Poppelsdorf see fig. 27)	6345	14 387	11 674

Expressed in numbers (tab. 14), in the AOI, the four main locations of the botanical garden only within the blue bin are responsible for about one third of the users, posts and userdays.

To relate the number of locations in a bin to the number of posts, hexbins can also be used. For this purpose, a variable can be assigned to the radius of the hexbins and the color shade (see WECKMÜLLER 2021e).

In this way, any combination between

- number of locations
- number of posts
- number of users
- number of userdays

or composed metrics such as posts per userdays can be represented.

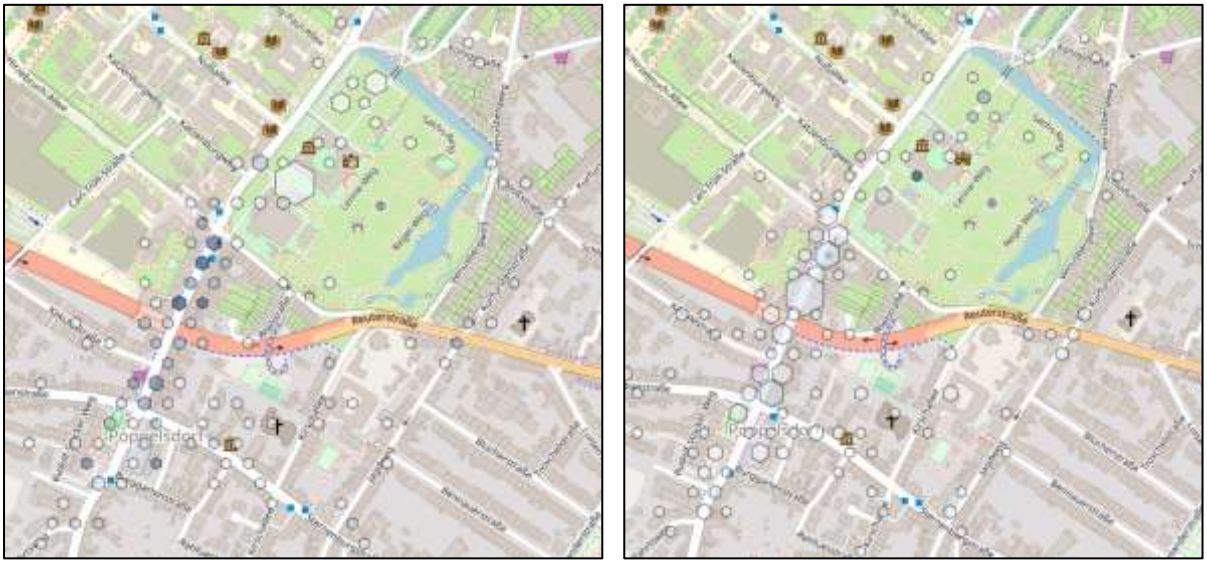


Figure 29: Two versions of hexagonal bins representing number of locations and number of posts for the variables number of posts (left: radius, right: color shade) and number of locations (left: color shade, right: radius).

Fig. 29 shows how the number of locations in a hexbin and the number of posts can be combined. On the left, the radius represents the number of posts and the blue shade the number of locations, on the right vice versa. Depending on particular analysis interest, these representations can easily be swapped. On the left, the actual importance on LBSM can be understood better, while on the right, the density of different POIs is in the foreground. Still, in both versions, the secondary information is always recognizable.

This process can now be repeated as often as desired for each custom AOI and varied with different combinations.

4.3.2 Temporal

For the temporal facet the same standard metrics can be displayed as for the spatial facet[22].

Table 15: Yearly distribution of users

Year	Users	Posts
2010	9	22
2011	93	520
2012	651	3164
2013	1607	6597
2014	3772	15 027
2015	8096	30 523
2016	15 846	63 579
2017	21 371	86 335
2018	22 763	95 793
2019	38 243	152 053
2020	40 353	194 547
2021	1502	3822
Total[23]	154 306	651 982

Table 16: Monthly distribution of users

Month	Users	Posts
Jan	12 031	47 810
Feb	9515	41 036
Mar	9224	40 301
Apr	12 612	51 953
May	11 137	48 005
Jun	11 055	46 140
Jul	12 481	50 508
Aug	15 215	61 215
Sep	13 665	59 257
Oct	14 872	64 768
Nov	13 857	61 463
Dec	18 642	79 526
Total[23]	154 306	651 982

Tab. 15 shows how the distribution of the different years behaves. For the year 2021, only January is included due to datamining at that time. Generally, due to the fact that random time periods or the data might have been sent by the respective API, one should be careful and not overinterpret these data. Tab. 15 is merely intended to show the increased number of posts year over year and the following overrepresentation of more recent posts in the dashboard.

[22] The metric of userdays for the temporal facet in lbsntransform and HLL-DB was not yet implemented at the time of writing.
[23] The total deviates from other tables as not all mined posts had timestamps.

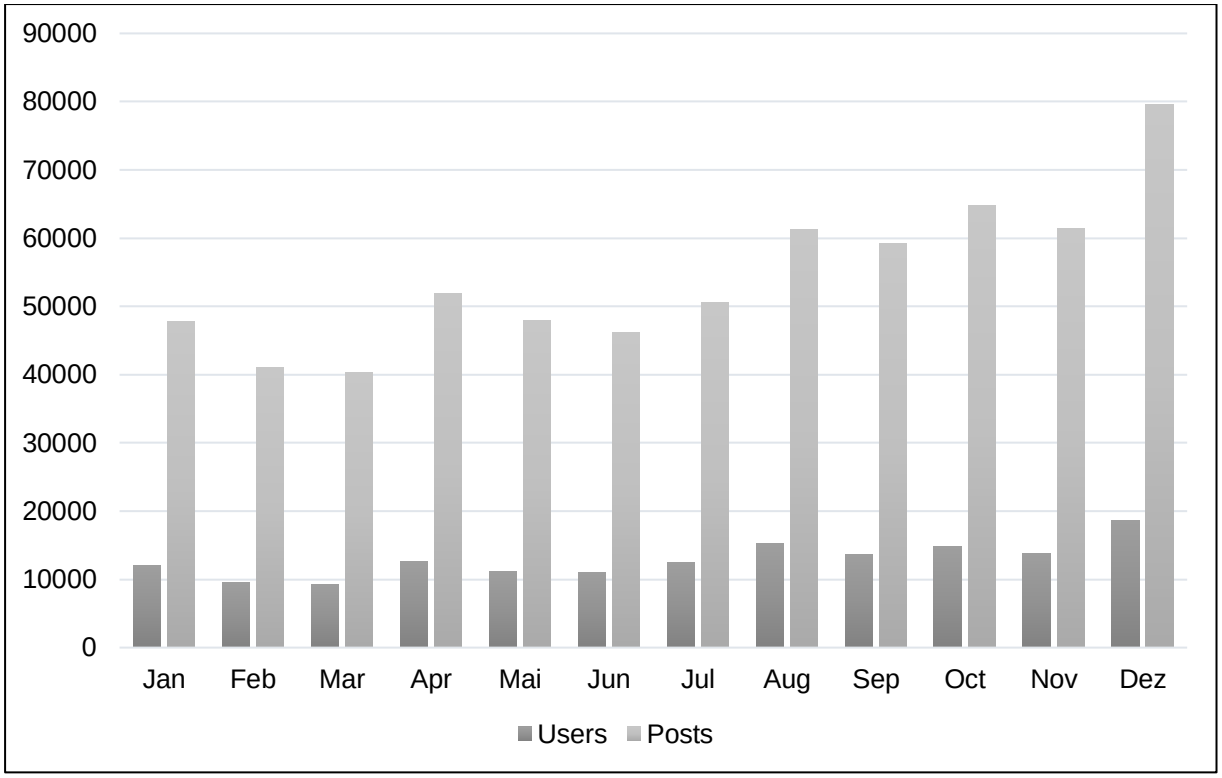


Figure 30: Monthly absolute users and posts for Bonn. Note: January is overrepresented.

Instead, tab. 16 and fig. 30 represent the monthly distribution where a careful interpretation is permissible, well-considering, that a slight bias for January is included in the data due to it being the only month analyzed for 2021.

Other temporal distributions like weekly, daily, hourly etc. were not yet fully implemented in lbsntransform and HLL-DB at the time of writing. These distributions could provide more insight than the rather coarse resolution of years and month. However, the APIs used here are, to a certain degree, non-transparent, making it impossible to fully rely on temporal distributions interpretatively.

Since on the one hand the temporal facet has not yet been implemented in the dashboard and on the other hand only becomes interesting in combination with other facets and therefore requires a higher level of complexity, it is dealt with in more detail in phase 2 (ch. 4.4.2). For future development a temporal selection option will be implemented in the frontend.

4.3.3 Thematic

The thematic facet deals with the content of the actual posts. Since the content is very diverse and can be divided into different atomic components, this facet offers a large scope for analysis that can only be partially covered here.

According to DUNKEL et al. (2021: n.p) the post caption can be divided into at least terms, emojis, hashtags and user mentions (tab. 9), the latter being assigned to the social facet.

Each of these atomic components can be examined spatially.

Table 17: Top 20 terms and hashtags in Bonn ordered by number of posts.

	Term	Users	Posts	Userdays	Category
1	bonn	50 942	20 7776	173 543	Location
2	germany	15 625	38 071	31 252	Location
3	love	11 650	25 822	24 086	Emotion
4	bonnstagram	1995	19 323	15 201	Photgraphy
5	igersbonn	1478	19 217	15 265	Photgraphy
6	instagood	5610	19 098	16 073	Photgraphy
7	rhein	7774	17 217	15 838	Location
8	deutschland	5879	16 119	13 014	Location
9	köln	5003	15 780	13 616	Location
10	happy	8141	15 153	15 065	Emotion
11	photography	5308	13 899	12 572	Photgraphy
12	art	5349	13 817	11 826	Culture
13	nature	5391	12 839	11 452	Nature
14	picoftheday	4227	12 833	11 509	Photgraphy
15	liebe	5576	11 968	11 155	Emotion
16	photooftheday	4219	11 518	10 055	Photgraphy
17	nrw	3235	11 305	9280	Location
18	food	3206	11 293	9892	
19	werbung	2464	10 463	9385	
20	beautiful	5704	10 449	9684	
	Hashtag				
1	bonn	20921	73503	62890	Location
2	germany	6916	14337	11983	Location
3	igersbonn	912	6850	5916	Photography
4	love	3014	6819	6350	Emotion
5	bonnstagram	1083	6726	5675	Photography
6	instagood	2135	6176	5288	Photography
7	deutschland	2411	5798	4775	Location
8	rhein	2497	5550	4816	Location
9	köln	1615	4478	3927	Location
10	picoftheday	1718	4283	3934	Photography
11	cherryblossom	2521	4135	3690	Nature
12	photography	1985	4090	3847	Photography
13	art	1353	3975	3050	Culture
14	beethoven	2847	3847	3578	Culture
15	altstadtbonn	605	3827	3279	Location
16	travel	2228	3726	3225	Location
17	fashion	1003	3612	3103	Culture
18	happy	1721	3572	3284	Emotion
19	nrw	1273	3461	2796	Location
20	photooftheday	1679	3444	3235	Photography

Tab. 17 shows the top 20 terms[24] and top 20 hashtags after exclusion of common stopwords in English and German (e.g. particle, article, filler words etc. see WECKMÜLLER 2021b: section "misc"). After researching these terms and hashtags on Instagram, they are divided here into categories partially following BURK, HUHN & WECKMÜLLER (2020: 12f) – in a short variant – of grounded theory "from the ground up" (GRUBER & HOLSTEIN 2014: 36, as cited in BURK, HUHN & WECKMÜLLER 2020), which should only serve to give an idea of which clusters appear most frequently.

These include, above all, locations on the spatial levels of country (Germany), region or state (North Rhine-Westphalia "nrw") and city ("Bonn", "Cologne"), photography-emphasized terms or hashtags of a photographic community on Instagram such as "bonnstagram", "igersbonn", "instagood", "photography", "picoftheday", "photooftheday" and positive emotions such as "love" in English and German as well as "happy".

Two hashtags are characteristic for Bonn. On the one hand, "cherryblossom" refers to Bonn's old town, which is home to several streets lined with cherry trees that are very popular among photographers and instagrammers when in bloom. On the other hand, "beethoven" refers to the 250th birthday of Ludwig van Beethoven, for whose anniversary a separate limited liability company "Beethoven Jubiläums GmbH" was founded for marketing purposes to coordinate various events around the core area of Bonn (BEETHOVEN JUBILÄUMS GMBH 2021: n.p.).

Table 18: Top 20 emojis in Bonn ordered by number of posts.

	Emoji[25]	Users	Posts	Userdays	Category
1	❤️	4157	8981	8446	Emotion
2	😍	3549	7884	7260	Emotion
3	🌸	3466	5723	5267	Nature
4	😊	2003	4097	3948	Emotion
5	DE	2730	4039	3672	Location
6	💗	1470	3192	2993	Emotion
7	🌞	1811	3023	3063	Nature
8	😉	1434	2875	2780	Emotion
9	✨	1580	2788	2700	
10	🎉	1312	2427	2363	
11	😂	1650	2366	2424	Emotion
12	🎄	1524	2231	2207	Culture
13	😋	1066	2167	2155	Emotion
14	😎	1232	2144	2088	Emotion
15	📸	1227	2036	1938	
16	🖤	1292	2009	1915	Emotion
17	😻	1013	1835	1742	Emotion
18	🔥	867	1791	1539	
19	🤗	840	1772	1679	Emotion
20	😁	995	1610	1573	Emotion

[24] In lbsntransform and HLL-DB all words of a caption fall under terms, including hashtags and user tags.
[25] The display of emojis varies on different platforms.

In addition to the terms and hashtags, tab. 18 shows the top 20 emojis that consistently express positive emotions. Here, the cherry blossom is the third-most used emoji indicating a possible reference to the cherry bloom in Bonn's old town. Among the very general emojis expressing some sort of happiness, the Christmas tree is more particular and could be an interesting object of research being possibly related to the Christmas market in Bonn as LBSM magnet. Further interpretation and an emoji comparison follow in ch. 4.4.3.

4.3.4 Social

Since no personal information such as age, gender, number of followers, friends, etc. was mined, a distinction is only made between different LBSNs, which generally have different user bases (see ch. 3.3.2).

As with the temporal facet, nothing can be derived without combining it with other facets, apart from the post distribution for this case study (tab. 11).

4.4 Phase 2 – Complex Queries for Bonn's Urban Green Spaces

Based on the discussion during the presentation of an earlier research project about UGS in Bonn (cf. BURK, HUHN & WECKMÜLLER 2020) on September 16, 2020 to the city of Bonn (see GEOGRAPHISCHES INSTITUT UNIVERSITÄT BONN 2020), some of the very concrete questions of the city of Bonn serve as an application-oriented focus for this use case. These questions can be summed up as follows:

How intensively are the UGS in Bonn used? When, by whom, why and for what activities?

While within the scope of this thesis not all of these aspects can be answered in detail, the focus is put on the spatial and thematic facet.

UGS can be defined as "public and private open spaces in urban areas, primarily covered by vegetation, which are directly (e.g. active or passive recreation) or indirectly (e.g. positive influence on the urban environment) available for the users" (HAQ 2011: 601, as cited in BURK, HUHN & WECKMÜLLER 2020: 2) e.g. "parks and reserves, sporting fields, riparian areas like stream and river banks, greenways and trails, community gardens, street trees, and nature conservation areas, as well as less conventional spaces such as green walls, green alleyways, and cemeteries" (WOLCH et al. 2014: 234, as cited in BURK, HUHN & WECKMÜLLER 2020: 2).

Thanks to the land use plan of the city of Bonn, the official geometries are available for all UGS, which may be used in the context of this work with the kind permission of the city of Bonn but may not be published as raw data themselves.

To answer the research question, complex database queries are used in the following, some of which are not yet implemented in the dashboard front end. However, most of the queries can be implemented with manageable effort, which is the subject of a future pilot study in cooperation with a municipality.

In addition, in this section all queries are combinations of different facets. They are assigned to the dominant facet or to the facet that is to be emphasized.

A selection of SQL queries is provided in the supplementary GitHub repository (see WECKMÜLLER 2021b).

4.4.1 Spatial

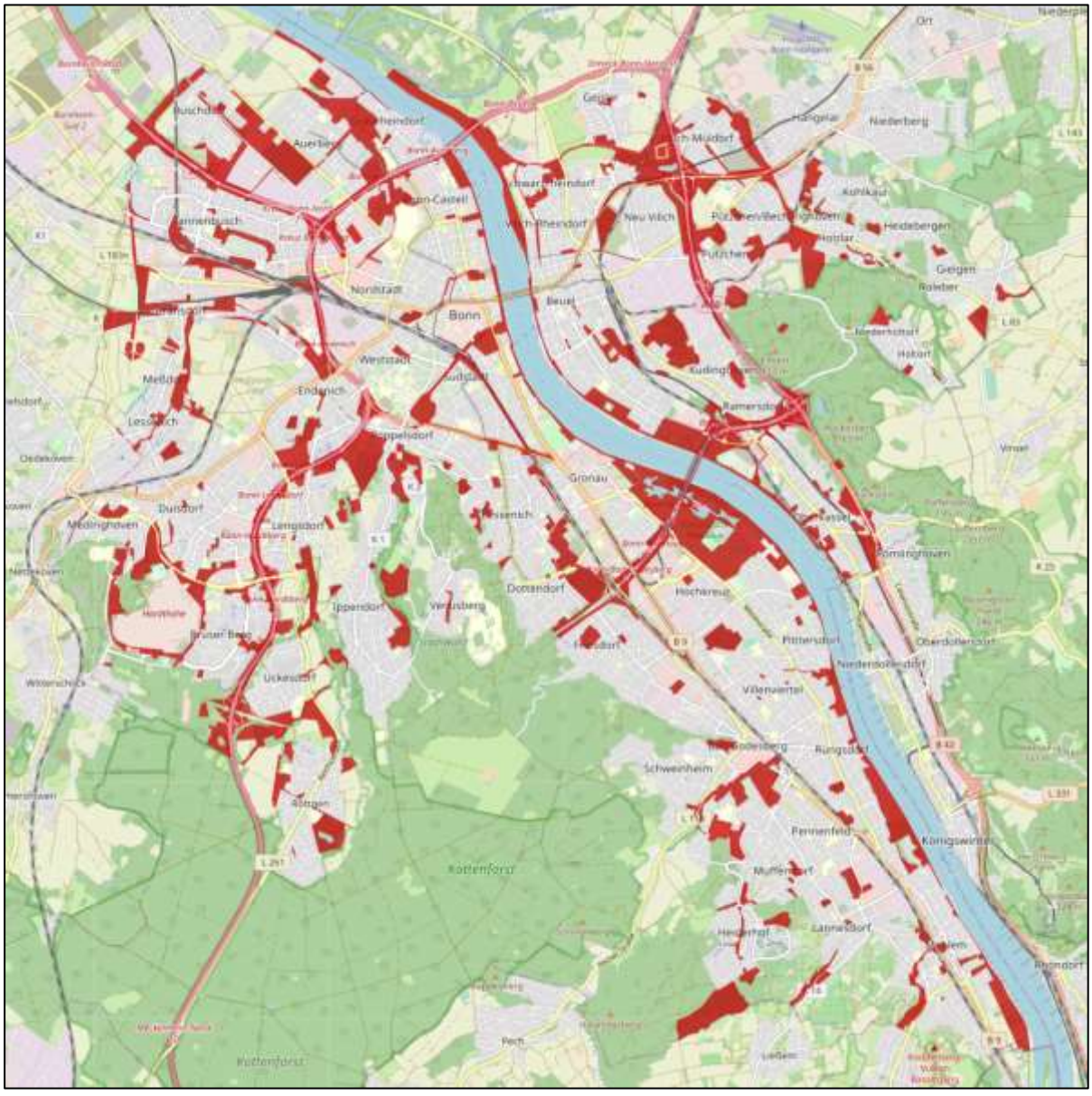


Figure 31: UGS of Bonn (red).

The land-use plan of the city of Bonn shows all areas of the category “green spaces” (DE: “Grünfläche”) in the dashboard after corresponding selection (fig.31). On the map, it can be seen that the areas are scattered practically over the entire city area. The “Kottenforst”, i.e. the forest area in the southwest as well as a part of the “Siebengebirge” in the northeast do not fall under the category “green spaces” but are listed as “forestry area” (DE: “Fläche für die Forstwirtschaft”). Therefore, it is not included here in the AOI. For all UGS together the metrics count 36 209 users, 92 238 posts and 75 698 userdays.

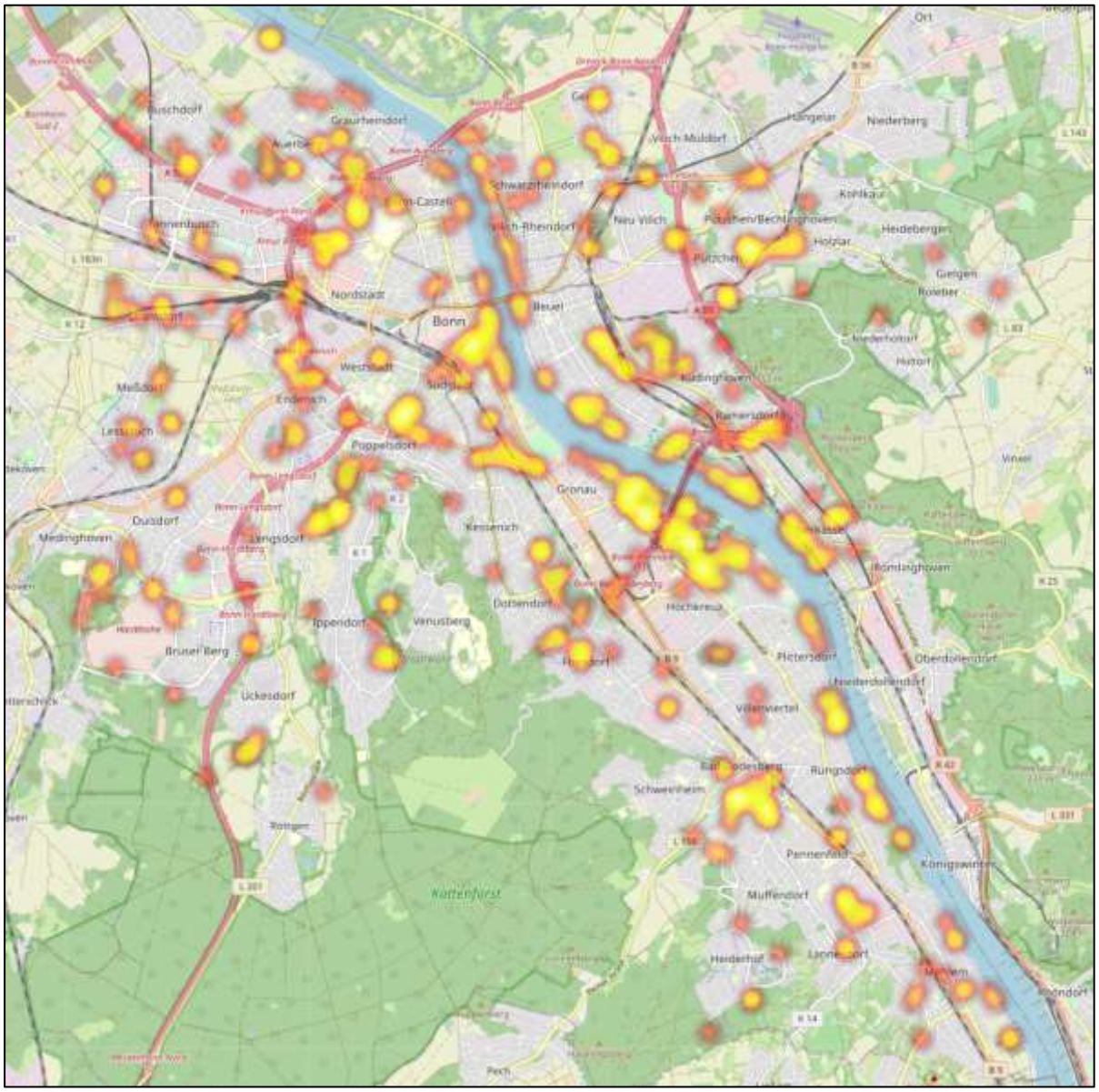


Figure 32: Heatmap for UGS in Bonn (yellow/orange).

Fig. 32 shows a heatmap similar to fig. 22 only for the UGS in Bonn. Some clusters can be recognized but are not as clear as before.

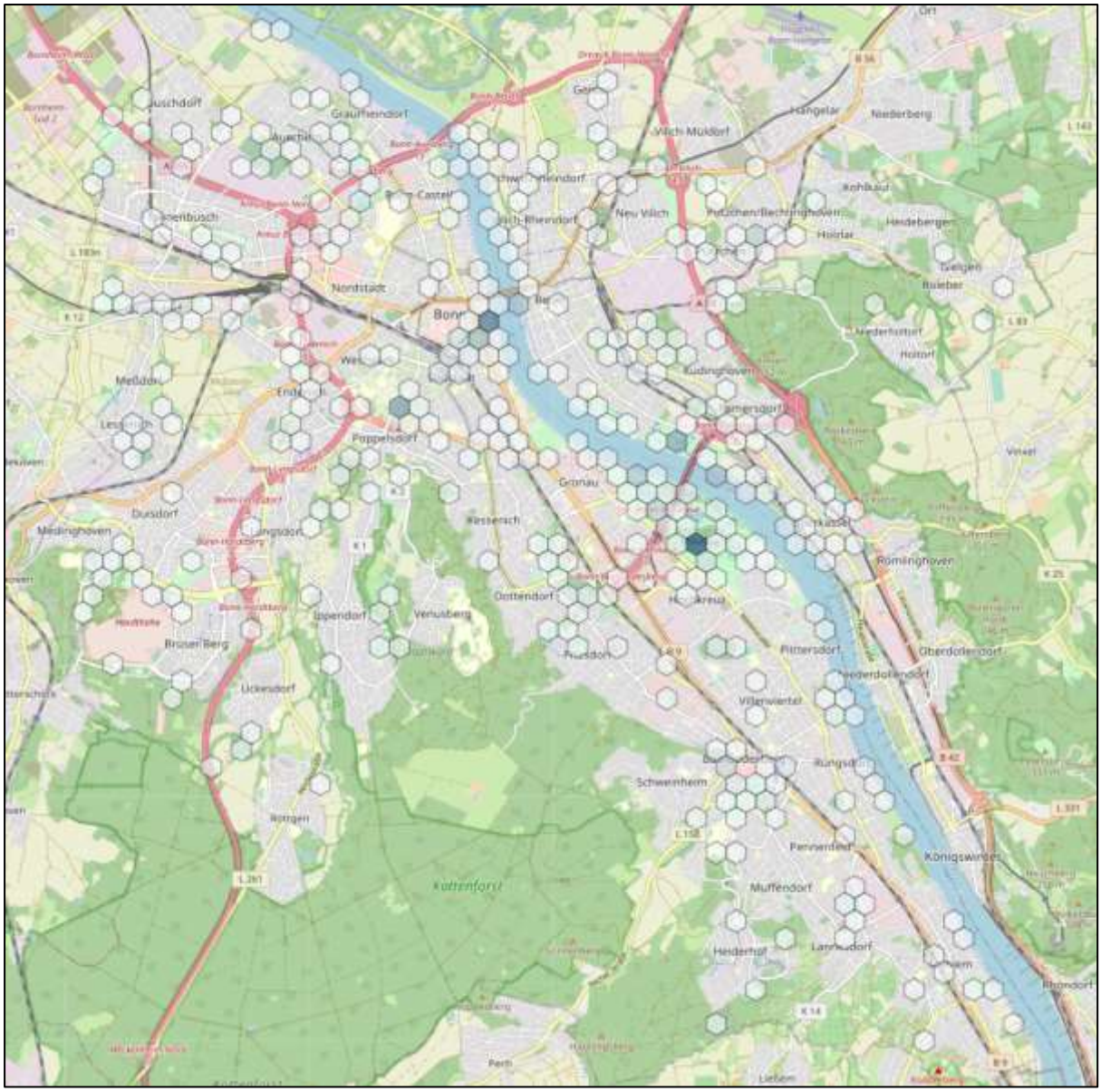


Figure 33: Hexbins for UGS of Bonn (blue).

Upon closer inspection by same-sized hexbins, hotspots and clusters become more apparent. Based on the number of posts, it can be quickly seen that certain UGS stand out strongly from other UGS in terms of quantity.

Figure 34: Selected UGS clusters.

When zooming in, the locations in the botanical garden, on the banks of the Rhine close to the center and in the Rhine meadows (DE: “Rheinauen”) can be easily recognized.

Aggregated to the individual UGS, a simple choropleth map could be generated. As this feature has not yet been implemented in the dashboard prototype, the following choropleth maps were created with the software QGIS instead. These maps should emulate the real appearance in the dashboard once implemented and are based on the same OpenStreetMap background layer.

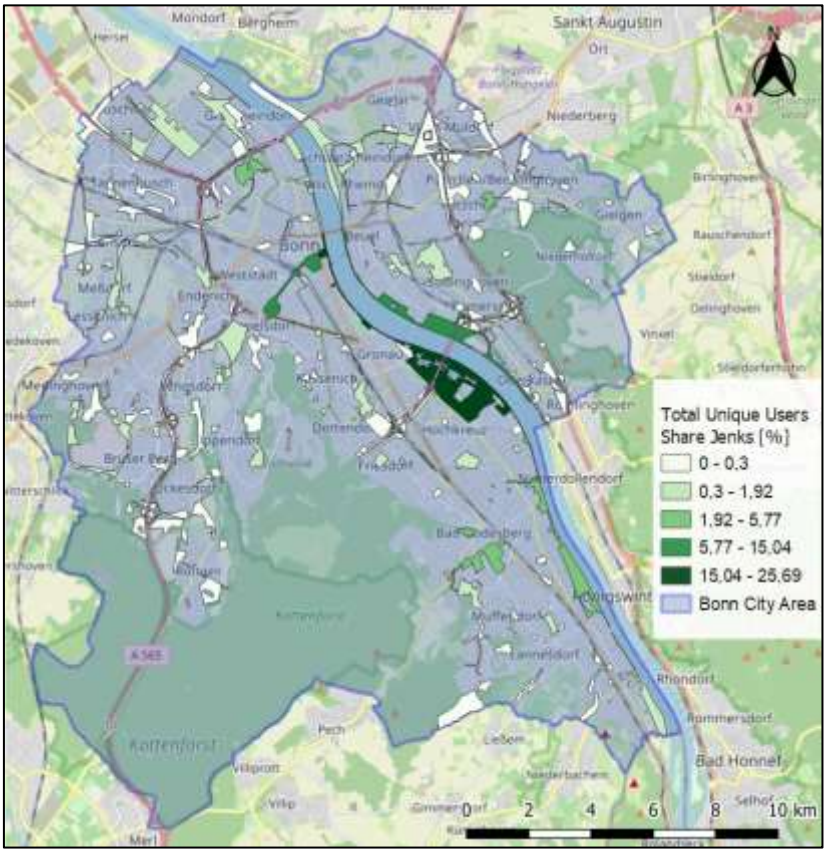


Figure 35: Total unique users share (jenks) for UGS in Bonn.

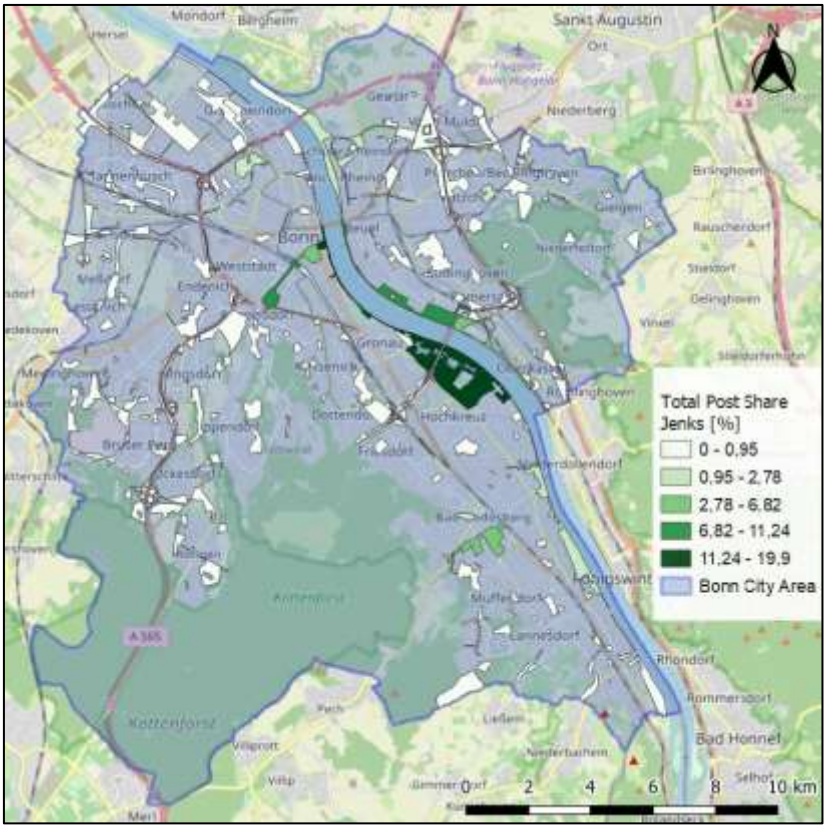


Figure 36: Total post share (jenks) for UGS in Bonn.

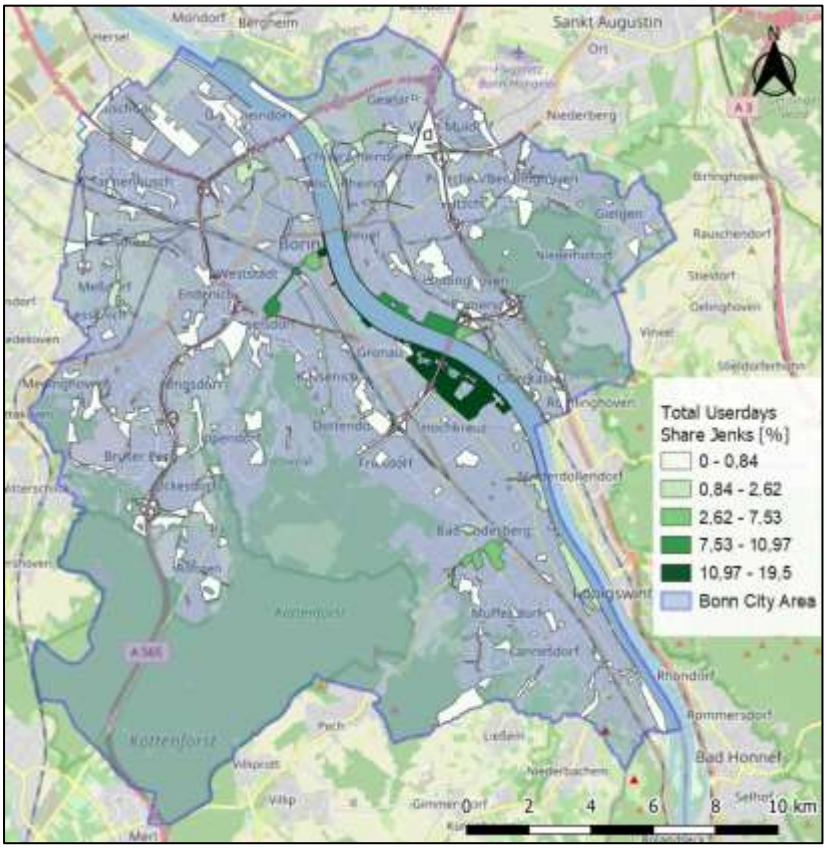


Figure 37: Total userdays share (jenks) for UGS in Bonn.

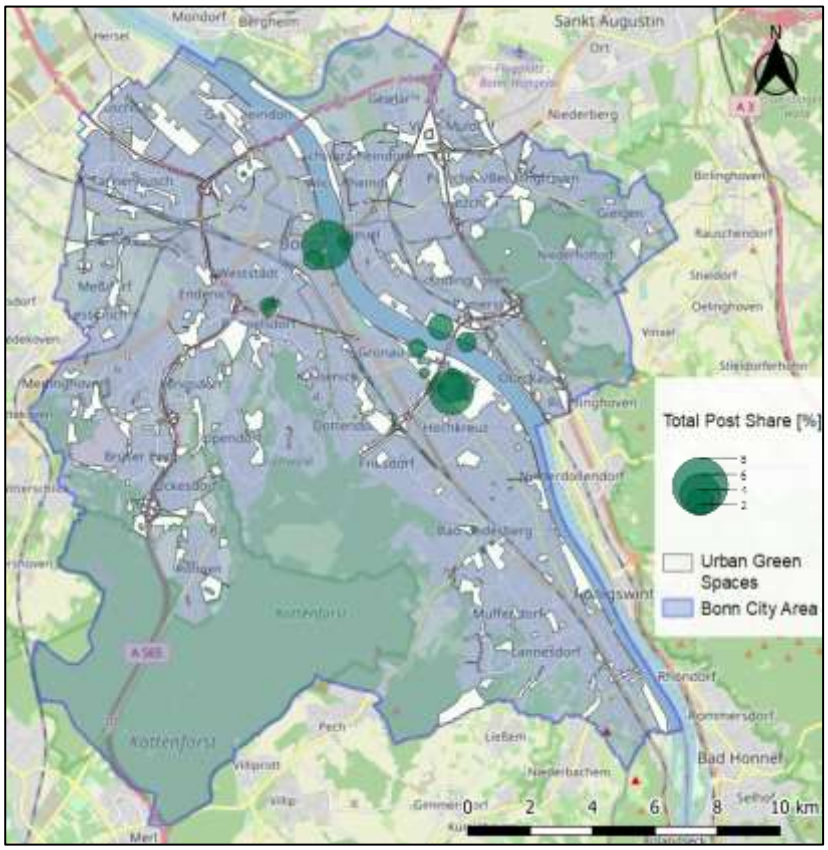


Figure 38: Top 20 locations total post share for UGS in Bonn

Table 19: Top 20 UGS locations ordered by number of posts.

	Location	Users	Posts	Userdays	UGS
1	241462299	3907	6057	5255	Rhine bank
2	218658046	2908	5683	4451	Rheinaue
3	214992500	2501	3913	3555	Rheinaue
4	278605025928845	2648	3466	3412	Rhine bank
5	188139351697395	2093	2839	2656	Rhine bank
6	648779	1768	2711	2487	Hofgarten
7	314155911	1501	2681	2204	Rhine bank
8	291797571	1390	2598	2187	Poppelsdorfer Allee
9	404804466543428	1249	2447	2088	Rhine bank
10	238010782	1236	1833	1584	Rheinaue
11	251543284	937	1567	1303	Poppelsdorfer Allee
12	ePRkT4tTUru5KjyCBA	102	1503	275	Poppelsdorfer Allee
13	236251931	588	1330	1129	Bad Godesberg
14	229774289	743	1301	1053	Bad Godesberg
15	230873913	892	1286	1115	Poppelsdorfer Allee
16	215416974	254	1271	914	Rheinaue
17	236292751	653	1267	1152	Sport Park Nord
18	113589042524900	804	1121	1054	Rhine bank
19	A.rtg7NXUbyYIRg	27	1098	47	Bad Godesberg
20	Uvujcu1XVrp6hVQ	237	1093	456	Rheinaue

Fig. 35-38 show how such choropleth maps based on different metrics could look like. In addition, tab. 19 provides an overview of the top 20 locations, sorted by the number of posts, which are visualized in fig. 38. The majority of these locations are Instagram locations; only locations 12, 19 and 20 are Flickr locations, which can be queried online in the Flickr API (see FLICKR 2021d) but in some cases might no longer exist.

Short Summary

The spatial attractiveness of UGS reflected on LBSN is very unevenly distributed. Depending on the metric, certain focal points can be identified. The eastern and western Rhine banks with the Rhine meadows (DE: "Rheinauen") are by far the most frequented LBSM locations. However, the UGS close to the center, such as the "Poppelsdorfer Allee" with the meadow "Schlosswiese" and the botanical garden as well as the "Hofgarten", enjoy also great LBSN popularity.

While the areas on the western bank of the Rhine, including the "Rheinaue", have a share of between 11 and 20% of the total post share, 15-26% of all unique users in Bonn have visited these areas and posted something there. Here, too, the main focus can be seen in the "Rheinaue" and generally on the Rhine. The top 20 most frequently selected post locations for can also be located within these areas. The trend can also be confirmed in the proportion of all userdays, i.e. the proportion of days on which users have posted something. The western bank of the Rhine alone is responsible for this proportion with approx. 11-20%. Individual locations as represented in tab. 19 can cause a certain bias if there are several highly frequented but not most frequented locations in one place, which virtually share the total amount of post. Proportionally, enormous focal points can be recognized, however alternatively, hexbins are preferable to such individual listings to avoid this phenomenon.

Derived from the focal points, the question how the LBSM distribution looks like cannot be answered absolutely. Instead, from the POIs two hypothetical LBSM attraction factors can be derived, the proximity to either the Rhine or university sites. However, these hypotheses cannot be further discussed here.

While the number of users (fig. 36) and the post behavior (fig. 37) show a high average attraction, based on the userdays (fig. 38) it becomes clear that the smaller, more peripheral areas are also used – quantitively significantly less, but they might play an important role for local residents. While this thesis is not intended to qualitatively evaluate UGS, it should be emphasized once again that only a section of reality is presented here and no planning measures should be derived on the basis of these one-sided results.

4.4.2 Temporal

At this point, as with all other facets, theoretically many different facet combinations can be analyzed. However, due to the fact that the HLL algorithm produces large errors for intersections with HLL sets of very different sizes, these temporal intersections are too error-prone which is illustrated here.

There were 92 238 posts in the whole investigated area of all UGS with an error rate of 3-5% resulting in the worst case to an error of +-4611 posts. In the month of July, there were 25 002 posts in March (as lowest month) with 3-5% error rate, which corresponds to +-1250 posts. Due to the fact that the error rate of the larger HLL set is dominant in a union, results that are below this error rate and only yield a few 1000 are practically unusable.

To evaluate a HLL set, the error rate can be derived from the errors of the individual HLL sets as follows. For $HLL_1 > HLL_2$:

$$\frac{HLL_1 * 0{,}05}{HLL_2} \tag{6}$$

Table 20: Monthly user and post distribution for UGS. Note: results are invalid!

Month	Users	Posts
Jan	3516	1316
Feb	2508	1819
Mar	1795	2219
Apr	5233	2127
May	4543	5927
Jun	1720	-2344
Jul	4175	4458
Aug	3766	1524
Sep	4488	790
Oct	1993	-2358
Nov	3567	789
Dec	4224	700
Total	41 528	16 967

Tab. 20 shows that the intersection does not yield any useful results, since on the one hand negative values occur (marked in gray) and on the other hand all values are below 6000 being close to earlier mentioned worst case error of +-4611 posts.

Instead of working with HLL intersections, it rather would be expedient to create certain aggregates as illustrated in ch. 3.2.7, e.g. from month, day of week, hour or similar and coordinates.

At the time of writing this has not been implemented yet in lbsntransform and HLL-DB but has been added by now.

This case study favors the other facets here and leaves the temporal as well as the social facet as subject to further research and development.

Short Summary

Due to the limitations of the HLL algorithm and its drawbacks with respect to intersections, no temporal conclusion can be drawn without the formation of aggregates when reading in the data.

4.4.3 Thematic

4.4.3.1 General Overview

For the thematic facet, thanks to the aggregation of term, hashtag or emoji, location and coordinate to a base combination during datastreaming, it is possible to proceed as previously described. Without these aggregates, the problem of HLL intersections with different HLL set sizes described before for the temporal facet would occur, which would make the results unusable.

First, the same statistics as in ch. 4.3.3 for all UGS are listed.

Table 21: Top 20 terms, hashtags and emojis for UGS ordered by number of posts..

	Term	Users	Posts	Userdays	Category
1	bonn	13 179	31 742	27 968	Location
2	rheinaue	3560	6700	6007	Location
3	germany	3269	5785	5280	Location
4	rhein	3309	5587	5198	Location
5	nature	1818	3783	3201	Nature
6	love	2450	3595	3527	Emotion
7	igersbonn	571	3461	2831	Photography
8	bonnstagram	596	2997	2408	Photography
9	instagood	1164	2630	2282	Photography
10	photography	1254	2460	2167	Photography
11	deutschland	1249	2441	2175	Location
12	summer	1646	2175	2170	Season
13	happy	1623	2068	1993	Emotion
14	photooftheday	1019	1993	1695	Photography
15	picoftheday	1014	1946	1775	Photography
16	meinbonn	182	1915	1436	
17	beautiful	1245	1880	1782	
18	sun	1279	1855	1704	Nature
19	friends	1307	1842	1691	
20	fun	1196	1820	1728	
	Hashtag				
1	bonn	4068	8648	7757	Location
2	rhein	1203	2228	1930	Location
3	germany	1198	1867	1655	Location
4	igersbonn	290	1428	1175	Photography
5	bonnstagram	293	1161	972	Photography
6	nature	441	947	785	Nature
7	deutschland	511	847	759	Location
8	hofgarten	538	801	764	Location

9	instagood	362	800	763	Photography
10	meinbonn	104	731	560	
11	bestofbonn	110	699	557	
12	photography	418	657	592	Photography
13	unibonn	361	634	591	
14	love	440	607	581	Emotion
15	beuel	247	600	535	Location
16	sunset	316	579	538	Nature
17	rheinufer	326	567	542	Location
18	autumn	353	555	503	Season
19	travel	343	535	453	
20	poppelsdorf	357	535	503	Location
	Emoji				
1	❤	511	634	632	Emotion
2	🌞	493	630	620	Nature
3	😍	423	585	583	Emotion
4	DE	361	464	415	Location
5	😊	270	347	341	Emotion
6	🌞	204	306	298	Nature
7	🍂	230	268	275	Nature
8	😎	196	255	256	Emotion
9	🍁	233	254	251	Nature
10	📷	205	245	242	
11	✨	197	230	223	
12	🖤	203	229	227	Emotion
13	💕	170	229	219	Emotion
14	😉	164	206	202	Emotion
15	🌸	181	199	195	Nature
16	😄	180	198	197	Emotion
17	🌳	105	197	191	Nature
18	💋	75	179	156	Emotion
19	📷	114	169	157	
20	💛	111	167	164	Emotion

Terms, hashtags and emojis (tab.21) for the UGS are pretty similar to the overall usage in Bonn. However, slight but important differences can be recognized. The importance of the Rhine for UGS according to LBSM is quite high, with “Rhein” as fourth most important term and second most important hashtag in comparison to the overall seven most important term and eight most important hashtags. Additionally, seasonal terms and hashtags such as summer and autumn as well as emojis (🍂, 🍁, 🌞, 🌞) appear to play a major role in UGS whereas cultural terms and hashtags apart from photography do not appear. The pink cherry blossom emoji on rank three for overall Bonn likely might be correlated to the cherry blossom (🌸) whereas the particular street is not included in UGS. However, it appears for UGS on rank 15. The regular tree emoji (🌳) is the 17th most popular emoji whereas for allover Bonn it is not included, but instead the Christmas tree emoji (🎄), probably linked to the general high impact of the Bonn Christmas market.

The description and interpretation do not claim to be complete but rather show the high potential for different purposes.

4.4.3.2 Targeted Queries

Thematically targeted queries with spatial UGS filter can be executed for very different purposes. For example, if the city wants to know which UGS are most used for the sports “Fußball, soccer, basketball, volleyball”[26], the results can be visualized quickly (fig. 39).

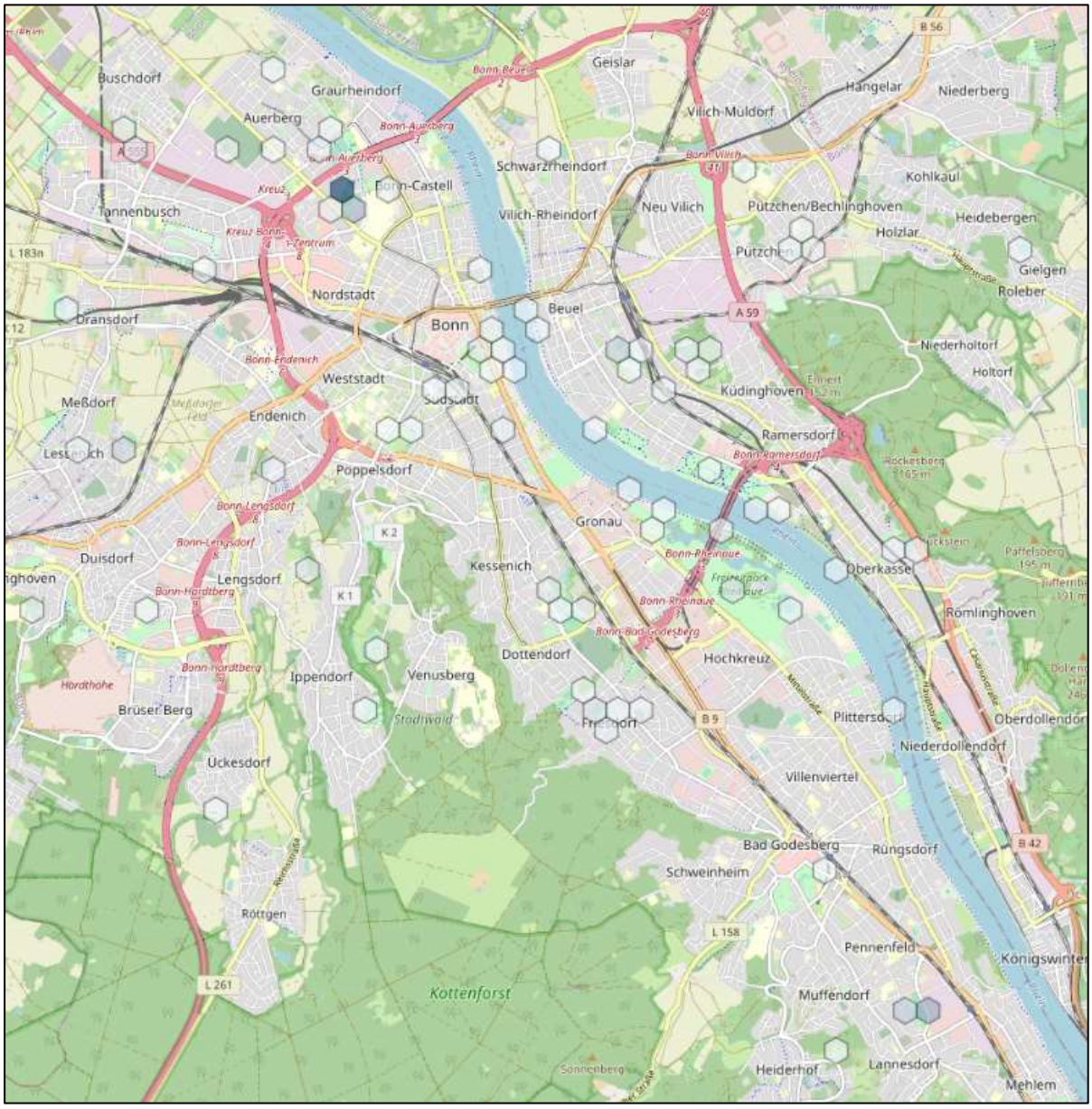


Figure 39: Hexbins for “Fußball, soccer, basketball, volleyball” in UGS.

[26] This was a literal question of the city of Bonn. Upper and lower case is ignored. As already mentioned, the query should be performed in English and German.

Taking a closer look at the sites with the most posts, it is noticeable that the city of Bonn considers sports fields and stadiums as UGS (fig. 40). For this reason, the most frequented places are obviously at sports fields, where LBSM posts are necessarily concentrated.

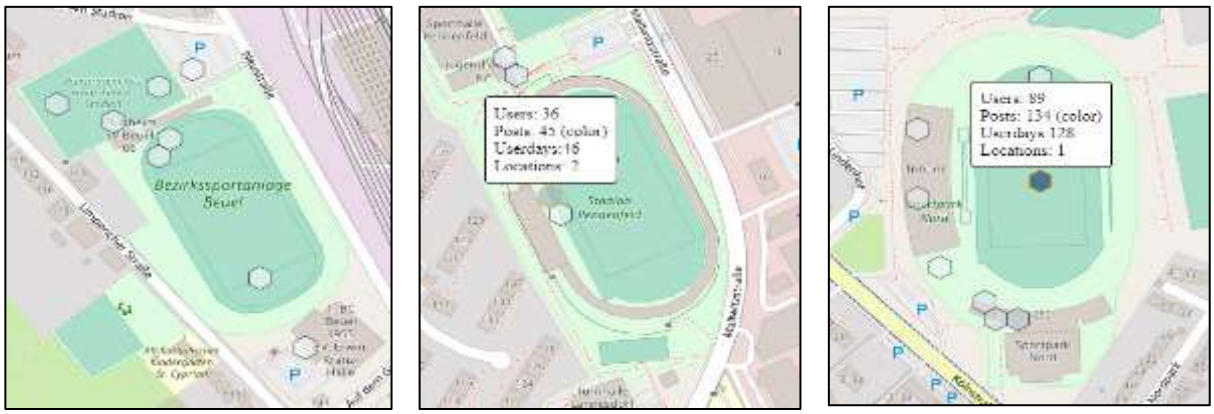


Figure 40: Sports fields selection in UGS.

4.4.3.2.1 A citizen-city scenario

From a citizen-city perspective for planning purposes, the dashboard could, on the one hand, provide useful information for citizens looking e.g. for a ping-pong table in a UGS.

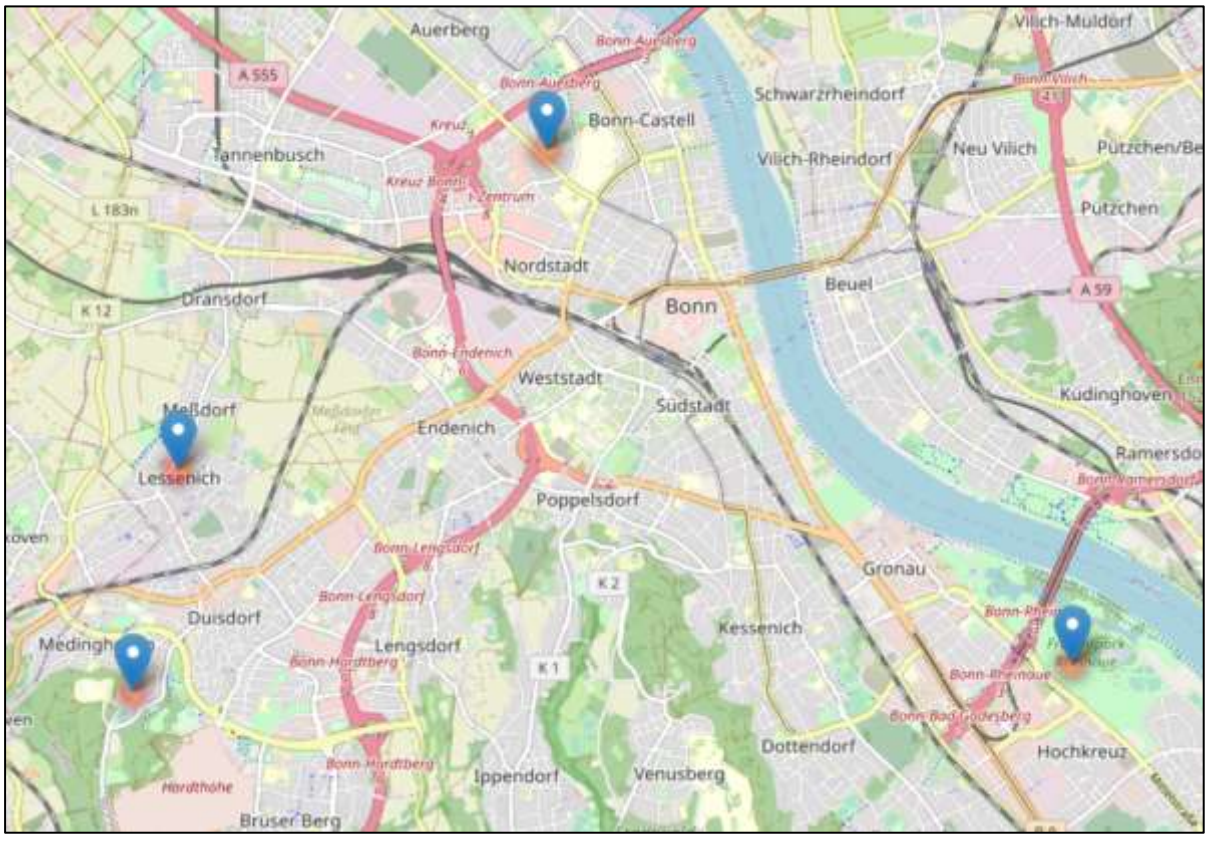


Figure 41: Locations and heatmap for query "ping-pong, pingpong, table tennis, tabletennis, tischtennis".

For this search, it quickly becomes apparent that only four different locations are found (fig. 41), which may possibly represent a deficit considering the size of Bonn with more than 330 000 inhabitants (BUNDESSTADT BONN STATISTIKSTELLE 2021: 4). On the basis of this result, a citizen can now on the one hand use this information for private purposes, albeit initially on suspicion, since just because a post is localized at a certain location, it does not automatically imply the presence of a ping-pong table.

On the other hand, the citizen could also turn to the city and complain about the lack of infrastructure. The city, in turn, could – especially if little to no data are available on such special topics – immediately check whether the complaint is justified and either provide the citizen with information about locations or compensate for the deficit.

4.4.3.2.2 A smart city scenario

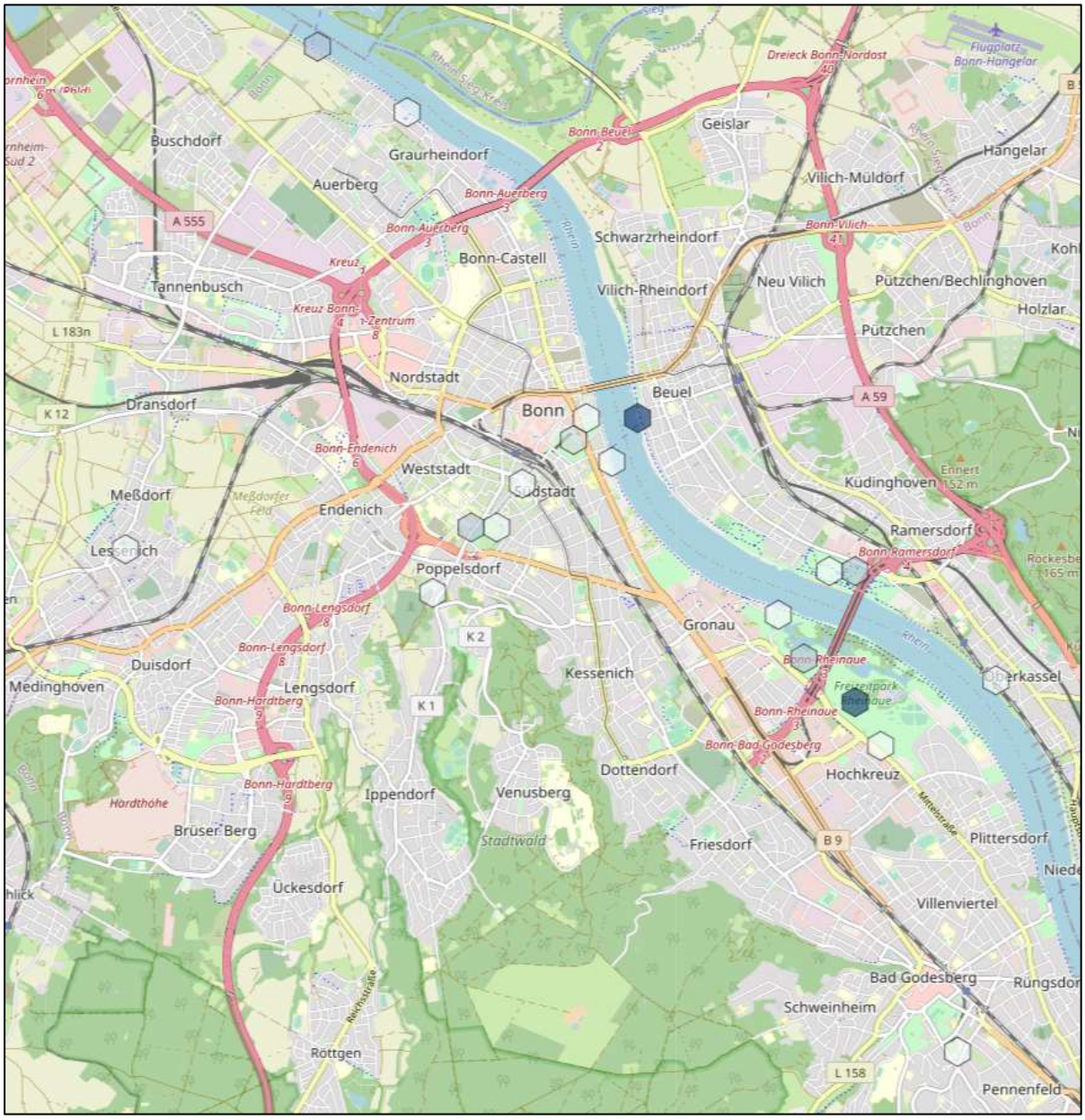


Figure 42: Query result for "Müll, garbage, trash, waste" in UGS.

The city could also easily use the dashboard to address issues like pollution or trash problems with the query: “Müll, garbage, trash, waste” (fig. 42). Since there is generally a positivity bias on SM, especially among adolescents (SCHREURS & VANDENBOSCH 2020: 2), proportionally fewer negative posts are to be expected. Nevertheless, the test for the Bonn data show that litter problems in UGS seem to occur mainly on the Beuel Rhine bank as well as in the “Große Blumenwiese” in the Rhine meadows (“Rheinauen”).

At this point, it must be emphasized once again that just because certain posts contain certain keywords, the reality at the location may look different, among other things due to the linguistic use of the terms in other, initially unexpected contexts (e.g. “Waste of time” or “trash party”). Nevertheless, a city can get a rough indication of where it might be necessary to check whether there is a garbage problem at a certain location. Thinking of a smart city approach this might even be automatized to a certain degree by sending for example a message to the respective office if a certain threshold of posts was registered, either relatively to the total amount of posts or absolutely.

4.4.3.2.3 City marketing

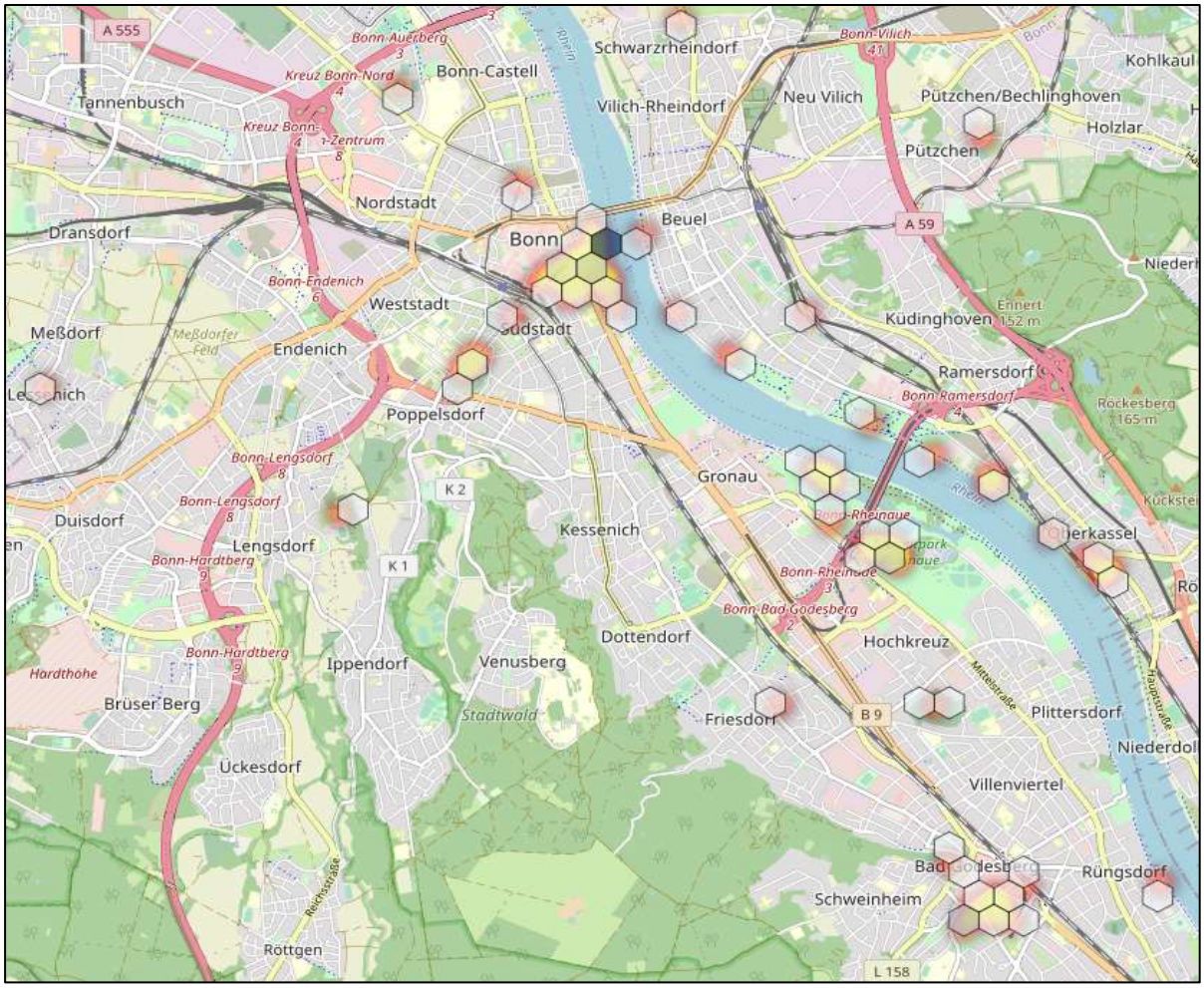


Figure 43: Query result for “beethoven” in UGS.

For marketing purposes, it would also be possible to analyze how successful certain advertising campaigns of the city are. The “Beethoven Year” with the large-scale advertising campaign, for example, could be analyzed and checked in which UGS the term “beethoven” appeared most frequently. This feedback is important for the city in order to target the expensive marketing campaigns to the right audience and eventually adjust the strategy if their benchmark is not reached.

On the basis of these data, a classic geomarketing approach for spatial SM targeting is thinkable also with other AOIs but is not further discussed here.

4.4.4 Social

As mentioned before for the temporal facet, the HLL sets should be as equal of size as possible for intersections.

This is however not the case for the social facet, so that it is not possible to work well without the base combination of atomic components when reading in the data.

Table 22: UGS metrics by LBSN. Note: results are invalid!

LBSN	Users	Posts	Userdays
Instagram	35 559	88 260	74 590
Flickr	335	11 893	846
Twitter	-293	1848	-2078

This can be explained well using the example of tab. 22. The corresponding intersection provides seemingly valid results, however they are afflicted with such a large error that they are practically unusable. For example, the result of 88 260 Instagram posts in the UGS is afflicted with an error of at least 26%, which can be derived from formula eq. 6:

$$\frac{478\,862 * 0.05}{92\,238} = 26\%$$

Again, further aggregation for base combinations during data streaming is proposed here.

4.5 Case Study Result Summary

The findings of the case study are briefly summarized here with a focus on the spatial and thematic facets.

4.5.1 Bonn Post Behavior

In general, postal behavior is strongly dependent on the spatial component. Heavily frequented districts such as the city center and Beuel, Poppelsdorf, Gronau, but also the “Rheinaue” as UGS and Bad Godesberg can be identified as LBSM magnets. While a wide range of gastronomy, stores, education and

culture institutions can be easily identified for the first districts with the help of OSM, the "Rheinaue" is not attractive due to its different locations and such a diversified offer, but rather as UGS per se – an attractivity that needs to be further investigated[27].

Since the majority of the sample data comes from predominantly photo-centric LBSNs such as Instagram and Flickr, it can generally be assumed that a certain visual bias prevails here, i.e. that presumably such locations which are visually more appealing than others receive more LBSM interaction, even if less aesthetic locations may be more popular in reality. What effect this bias may have cannot be discussed further here.

This photo affinity can also be seen in the numerous photo- and photo community-related top 20 terms and hashtags. The thematic facet can be divided into different clusters for the whole of Bonn covering the majority of posts. These clusters include primarily location terms or hashtags (which is a common practice in SM), photography (as expected), emotions and culture.

The most frequent emojis include strikingly many cherry blossom emojis (🌸) and Christmas tree emojis (🎄) underlining the importance of the cherry blossom and Christmas in Bonn in a certain way. However, the qualitative meaning cannot be read off without appropriate interpretation, which is stressed here in reference to BOYD & CRAWFORD (2012: 666f) once again.

4.5.2 Bonn Urban Green Spaces Post Behavior

A similar trend can be observed for the UGS in Bonn as for Bonn as a whole. The UGS show strong differences in use. The LBSM focal points are mainly located in the "Rheinauen", the central "Hofgarten" and the "Poppelsdorfer Allee" with the "Schlosswiese". Together, these three UGS form the LBSM focal point.

In general, the enormous importance of two factors can be derived for the UGS. On the one hand, the relation to the Rhine is of crucial importance for the post behavior on LBSM, which can be seen in the river banks and the "Rheinaue". On the other hand, the likewise highly frequented UGS "Hofgarten" and "Poppelsdorfer Allee" with "Schlosswiese" are characterized primarily by their centrality.

While the thematic facet largely follows the general trend for Bonn as a whole, subtle differences can be discerned, primarily in a clear reference to the seasons in the UGS. The focus on the Rhine and the "Rheinaue" can also be seen in the terms and hashtags.

[27] See BURK, HUHN & WECKMÜLLER (2020) for a cultural ecosystem services-based approach for UGS analysis in Bonn.

4.6 Phase 3 – Outlook

Apart from the basic functions for all facets, especially for the missing temporal and social facet, which are to be implemented as aggregates in lbsntransform and HLL-DB and subsequently in the dashboard frontend, several ideas are proposed here on how the dashboard could further evolve towards SG.

As best-practice city dashboards such as the Dublin Dashboard (see MAYNOOTH UNIVERSITY 2021) show, such a dashboard can serve as a citizen information hub. However, the focus for the LBSN dashboard should not be on general and quite trivial information such as e.g. the weather. While this information might certainly be useful but not exclusive, the LBSN dashboard should be used specifically as a democratic tool and platform.

To promote MSGG, for example, citizen petitions and participation procedures could be embedded transparently in the dashboard. An additional layer could display all petitions with their corresponding areas (similar to the land use plan, fig. 18), so that all queries can be performed by citizens and planners alike, just like in the case study.

In a further step, this idea could be expanded to create an interaction platform through which not only information is transferred from citizens and planners, but also information from citizens directly to planners and vice versa. Public comment functions linked to residential status could encourage local people to participate more easily than is the case with conventional often rigid participation procedures.

The LBSN dashboard could thus be expanded into a general participation and information hub, which would be at the heart of a SG approach. To this end, additional information could be gradually incorporated, such as official statistics or demographic data.

The practical advantage of the prototype developed here is that additional layers and query options can be easily added to the existing infrastructure. For example, using the functions already available, a layer with the current, ongoing public participation procedures could be incorporated in much the same way as the Bonn land use plan.

The basic prerequisite for such an infrastructure is a strong awareness of open data, for which the foundation has been laid in Bonn with the city's open data portal (BUNDESSTADT BONN 2021d). As part of the smart city initiative of the city of Bonn (BUNDESSTADT BONN 2021a), not only data from the platform, such as parking garage utilization, but also new transport projects, such as a new special bus route, could be recorded directly on the map, which, for example, could serve certain hotspots at peak times, which can be analyzed by the LSBN dashboard.

In a further step, the already existing API, through which the data are transmitted to the dashboard, could be further developed and released for public use. In this way for example, a smart city cleaning

system would also be conceivable, by calculating the route and sending a cleaning team to polluted locations if necessary.

For research, a combination of the quantitative methods applied in the dashboard and qualitative methods shall be proposed here, to gain more profound insight. Such a mixed methods approach as proposed e.g. by KUCKARTZ (see 2014) could be useful when investigating certain spatial clusters in order to evaluate why certain locations or areas have such a high LBSM attraction. The dashboard can deliver the starting points for qualitative research and support it along the way when applying e.g. a hermeneutic approach as described in ch. 3.1 (tab. 3).

The possibilities for extending the dashboard are practically unlimited and need further research.

5 Discussion

At this point, a critical evaluation of the performance, advantages and disadvantages in practice follows on the basis of the previous case study. It is followed by an assessment of the extent to which the dashboard can be a practical tool for this purpose, based on the MSGG definition and criteria from ch. 2.1.4.

5.1 Privacy

Since the technical focus of this thesis is on a privacy-aware infrastructure, it is discussed at the outset on the basis of the insights gained from the case study.

Firstly, with HLL-DB (see DUNKEL & LÖCHNER 2021a), HLL provides a solid basis on which to build by breaking down posts into atomic components such as terms, hashtags, etc. and creates a separate HLL set for each of these components.

At this point, it has become clear in the case study that, on the one hand, even small HLL sets with a set size <10 can be significant for gaining knowledge, especially for niche topics such as the example of ping-pong tables or garbage collection. At the same time, however, this is accompanied by certain losses in privacy, which may have to be evaluated much more critically in other contexts than in the example of table tennis. Here, too, it must be analyzed exactly what the dashboard is to be used for and what degree of possible privacy loss can be justified.

In addition, an unfiltered reading of all post captions bears the risk that plain names or direct user mentions (“@userABC”) appear here. This of course poses a risk to user privacy, which can be circumvented in various ways.

On the one hand, as explained in fig. 10, certain filter options can be set, i.e. white or black lists, which either allow or exclude certain terms. The former is to be preferred, because plain names or certain user

mentions can hardly be predicted. User mentions can be easily recognized and filtered out by the "@" symbol which is not the case for plain names.

For the dataset of this case study, it is questionable why a possible attacker would perform the query in the dashboard, which requires more effort, and not directly in public search engines. Nevertheless, this is an important problem area when using public data, especially since some posts that are delivered by APIs are not indexed by some LBSNs based on individual user settings. Here, the privacy violation would be considered particularly serious. Nevertheless, apart from the system-immanent privacy, additional methods such as cryptographic hashing, salt-key etc. need to provide strong protection against penetration.

In general, it should be emphasized once again here that the extent to which the privacy risks are acceptable must be weighed up and discussed in detail beforehand for each use case. In case of doubt – especially with regard to the pilot study – it is suggested to test phase 1 (ch. 4.3) in a non-public version first in order to exclude that attackers from outside can endanger the privacy of the users, even if this does not exclude that attackers could come from the internal team itself.

In general, however, it should be emphasized at this point that, despite the common assessment that "privacy is a side effect" of HLL (DUNKEL, LÖCHNER & BURGHARDT 2020: 3), it offers a strong basic protection, since the raw data are converted into probabilistic abstractions avoiding the full reconstruction of the original data set and this per se entails a privacy gain.

If, for example, a whitelist of "tree, baum, forest, wald, natur, nature" for nature-related terms and a minimum size of HLL sets were used, a very solid privacy-aware basis would be created when additionally applying cryptographic hashing and different salt-keys for each set. However, the methods need to be evaluated beforehand and discussed in detail.

Privacy Models

Applying the privacy model of XU et al. (2014: 1151), the instances of the data provider, i.e. the LBSN user, as well as the data collector, i.e. the different LBSN do not need to be discussed for this thesis. Instead, two instances, the data miner or rather the dashboard server as well as the decision maker are of interest.

As the dashboard offers plenty of possible settings, it is the most important privacy instance.

Here, with regard to the privacy dimensions of BARKER et al. (2009: 44) initially, the dashboard should be accessible only to the municipal office in order to test the settings and discuss it with the citizens before opening public access. At the final stage however, the dashboard aims to be publicly available, i.e. to the whole world (fig. 6). The scope or granularity (fig. 6) of the data usage can be initially very limited to certain topics of interest, i.e. presets for e.g. nature or UGS-related interests or the category

of sports. In this way certain purposes, e.g. commercial abuse can easily be prevented. The selection of presets highly depends on the interest of the municipality and the people and needs to be discussed beforehand.

However, assuming a successful and constructive municipality-citizen discourse and a dashboard with public access and high granularity, i.e. no thematical boundaries, a very high potential lies in the dashboard for different purposes and the public. At this stage, for privacy, there is no difference between the granularities, as the HLL sets have the same effect of a basic privacy for any set above a certain threshold.

Considering the dimension of retention (BARKER et al. 2009: 44), it is fully up to the municipality and the citizens to decide about the temporal scope of the HLL sets, e.g. if data for a year, a month or a week are used and to decide how long they are accessible. For privacy, it would be safest to delete HLL sets in a fixed interval in order to avoid the worst case of a successful attack on the server. In this unlikely case, the attacker could only get the most recent HLL sets but not the ones from the past.

With regard to the post facets (fig. 8), for the maximum of privacy as few facets as possible should be used. If certain facets are irrelevant to certain questions, they should be completely left out. In this way, a hypothetical attacker could gain even less insight.

To the instance of the decision maker technically the same core concepts apply as to the dashboard privacy. Generally, the derived information depends on the dashboard settings and is hence already limited which in turn, guarantees the minimum privacy level of the dashboard.

However, there is a fundamental difference in the two instances. While the dashboard server holds the real HLL sets, the decision maker (or any other citizen or institution) can only obtain the cardinalities. In this way, the information is – if certain minimum post thresholds are applied – harmless to the individual users' privacy.

Summing up the evaluation based on all privacy models, the dashboard offers plenty of options to use only the needed data and discard the rest. True to the motto "less is more", a careful evaluation based on the needs of the decision maker and citizens needs to be done. If all previously described precautions (ch. 3.2.8) are taken, there is a low risk of privacy breach. Additionally, as the dashboard must only use already public data, the privacy risk to an individual even in the worst case of a successful attack is very low.

5.2 Advantages and Disadvantages of HyperLogLog in Practice

As the case study and the dashboard frontend show, a solid full-stack web app can be developed based on HLL and LBSN-Structure with lbsntransform and HLL-DB, which is suitable for productive use.

The strength of HLL is generally in dealing with big data. With a small amount of data, HLL sets are error-prone, which is why in practical implementations, it is possible to switch between different modes to balance between privacy and accuracy. Such a switch is found e.g. in the PostgreSQL HLL-implementation of the company CITUS & CONTRIBUTORS (2021).

With HLL unions, HLL sets can be merged without loss, i.e. without growing errors, which is crucial for aggregated statistics. HLL intersections, on the other hand, exhibit major problems in practice due to the large error rate associated with HLL sets of different sizes. Thus, the case study has shown that HLL intersections are virtually unusable for the operations listed here. These can at most lead to a general indication dealing with error rate far beyond the usual 3-5% (DUNKEL, LÖCHNER & BURGHARDT 2020: 7).

This problem can be solved by creating aggregates of atomic components resulting in base combinations. The more these are aggregated, the more information is available in the dashboard. However, this also proportionally increases the privacy risk, which must be taken into account. For example, an aggregate of UID, coordinate and date can be extremely sensitive in the (still unlikely) event of a full penetration by an attacker including hash functions and salt keys that would severely compromise the user's privacy.

HLL is fast and memory-efficient and in combination with PostgreSQL as HLL-DB very performant. The practical advantage here is the use of SQL for targeted simple queries (see WECKMÜLLER 2021b: section "SQL").

The combination with LBSN-Structure has turned out to be extremely useful, since in such a way in one step, data of different LBSN can be merged into a DB and be queried immediately. The breakdown into atomic components, as it happens in the HLL-DB, is especially practical for the work with the thematic facet, as the HLL sets for different terms can be merged without loss.

However, certain profound analysis such as classic topic modeling or similar machine learning algorithms which work for raw data are prevented by HLL. This limitation must be considered for individual use cases.

5.3 Dashboard Implementation

The prototype developed here already covers the spatial and thematic facets, whereas the temporal and social facets still need to be extended in the backend in lbsntransform and HLL-DB.

However, the facets already implemented in the case study show the information potential that lies in a targeted use of LBSM data. There is no need for complex deep learning algorithms such as topic modeling to gain high resolution insight into the data, but instead very basic queries deliver very detailed

information and are sufficient for many purposes. Still, such sophisticated algorithms could still be developed in the future.

Spatial queries in particular make it possible to identify distributions and hotspots within a city. In this way, LBSM clusters can be discovered and used. Even if the results of the dashboard always require a corresponding, adequate interpretation, which presupposes a certain basic competence and an understanding of data, the dashboard can also be of considerable use to inexperienced users (laypersons) for simple questions, even without great expertise, if, for example, table tennis spots are simply being searched for. For such a trivial but potentially useful query, to the end user it does not matter, how many people or which age cohorts frequent the spots or whether some might be missing. The coordinate alone in combination with the simple tag "#pingpong" is enough to provide a clue of where to start the search for respective infrastructure.

The exemplary combination with the land use plan of the city of Bonn has shown how valuable the results of the dashboard are, if they are well-targeted. The distribution and types of use of the UGS, for example, can be easily displayed and evaluated.

Also, with regard to the problematic situation of social distancing caused by the COVID-19 pandemic, cities could assess where a particularly large number of people (mainly young people due to the platform demographics) come together and set appropriate incentives to prevent agglomerations in UGSs.

In absolute terms, the "Rheinauen", which are very large in comparison to other UGSs, already have the highest post number. However, if this number is set in relation to the UGS area, it emerges that the "Poppelsdorfer Allee", for example, with its relatively small area, has a significantly higher interaction rate on LBSM than the "Rheinaue". This fast is not surprising considering the centrality of "Poppelsdorfer Allee" and the relative remoteness of the "Rheinaue" southeast of the center.

The city could, for example, simply improve public transport services after a detailed examination of this hypothesis, so that the "Poppelsdorfer Allee" would be relieved and social distancing would be simplified – all based on the information from the dashboard.

With the previously described plugins, this information could be even better interwoven with other applications and democratic tools like citizen participation processes to ultimately govern smartly.

5.4 Municipal Smart Good Governance and the Dashboard

The working definition, which was established in ch. 2.1.4, is finally related to the dashboard and discussed here.

After the detailed treatment of privacy in the previous chapters and its practical meaning in the dashboard, it is first noticeable that it does not appear in the definition of MSGG. Privacy is so far reduced

to a complete niche existence in previous SG literature and has only recently started to gain momentum (e.g. JANGIRALA & CHAKRAVARAM 2021).

Thus, before delving into the actual performance of the dashboard for MSGG, after the detailed discussion in ch. 3.2.8, the importance of privacy should be acknowledged. After all, no matter how "smart" the framework is, it would not be able to help MSGG if it is harmful to LBSM users.

The extended definition shall therefore be:

> **MSGG is the capacity of improving municipal democratic decision- and policy-making processes and outcomes through the transparent and open usage of ICTs such as LBSM in order to increase citizen's quality of life and well-being without compromising the privacy of LBSM users.**

Beginning with the first part of the definition for a smart and well governing municipality, it is to be discussed how the dashboard could improve municipal democratic decision- and policy-making processes.

Through the LBSM structure with its facets, this can be answered quite concretely. The municipality can now answer questions for which there were simply no data before, thanks to the bundled access to LBSM data. As a framework, the spatial and temporal facets already offer precise bounding possibilities for coordinate and hourly up to yearly filters. These facets have already been partially incorporated in the prototype developed here and give an idea of the potential the dashboard would have in a fully developed version. Adding further facets through aggregates, very precise queries would be possible, delivering the basis for well-targeted decision-making.

Nevertheless, it should be emphasized here, just as in the previous chapters, that the results – with the exception of very trivial queries for simple needs – require appropriate interpretation, which presupposes a certain basic competence and understanding of the data. The data used in the dashboard always contain various biases and inaccuracies that are not immediately apparent from the numbers provided in the dashboard.

Rather, the suggestion for this is that the dashboard should not replace conventional methods such as citizen participation processes in the city districts, information campaigns, and communication platforms, but rather be used in a supportive manner. Especially for younger target groups, who are numerically more likely to be represented on SM, the dashboard can thus close a potential information gap and, if interpreted correctly, can be used in such a way that quality of life and well-being are actually increased.

The dashboard cannot replace political discourse either, but can only partially contribute to the factual basis. The actual negotiation processes must not be limited to simply reading off the dashboard, but should certainly take place between municipality and citizens.

At this point, the dashboard could run the risk of raising wrong expectations if regarded as a “salvation” tool by the municipality and used excessively and incorrectly. It is important to counteract this danger with open information and workshops so that the personnel who operate the dashboard not only learn how to use it, but also know about the initial data limitations and how to interpret the results.

Also, over an initial exclusive phase of the introduction of the dashboard, mistrust could arise if the dashboard is only used exclusively by the municipality. In this regard, the dashboard must be used transparently and openly over the long term so that it ultimately becomes a dashboard based on data from the citizens, hosted by the city for the citizens.

The extent to which quality of life and well-being are actually increased cannot yet be foreseen in this work. A good information base is the basic prerequisite, which of course must be used accordingly through certain mediation processes and translated into practical measures.

The definition of MSGG including its key aspects from the respective subconcepts (tab. 2) could be used to guide as well as evaluate the pilot study and propose appropriate adjustments if necessary.

For a theoretical guidance, the concepts of spatial equity and equality should be conceptualized further for an application-oriented usage. The discussion in this thesis is not sufficient to do so but rather offers a theoretical starting point for further research.

6 Conclusion

6.1 Practical Findings

The main insight gained in this work is, in short, that the combination of HLL, LBSN-Structure, a secure backend and the powerful frontend developed here provides an extremely solid privacy-aware basis that should be used for municipal decision-making.

The gain in privacy compared to the use of raw data is significant even in the privacy-poorest variant, although the strict definition of differential privacy cannot yet be fulfilled in the current version of the HLL-DB and lbsnstructure. At this point, further research is necessary to possibly create the final quantum leap from a privacy-aware to a privacy-preserving dashboard.

When it comes to concrete application, it is particularly important to weigh up the ethical aspects between the privacy of the user and the potential benefit for the general public in terms of increasing well-being and quality of life for each individual use case and scope of use. This cannot be defined here in a general way, but necessarily remains the subject of a lively, discourse-oriented democratic society. The

dashboard offers plenty of possible settings on a range from the already high minimum level of privacy (e.g. by aggregating base combinations with more than two elements) up to an almost-privacy-preserving level i.e. by using whitelists, different salt keys and no aggregates.

6.2 Case Study Findings

The case study for the Bonn area showed for the data sample of more than 650 000 posts that there are enormous differences in LBSM post behavior for different areas in Bonn.

Firstly, there exist individual hotspots in Bonn that attract most of LBSM attention. Depending on how the results are aggregated, interpreted and thereby summarized, it is already possible to see with a glance at the OSM base map what could be attractive about highly frequented spaces. For the formulation of such theses, the dashboard is perfectly suited. However, the dashboard alone cannot provide a conclusive insight. Rather, the theses still need to be tested also with conventional methods.

Such hypotheses were, for example, formulated in the study of UGS, i.e. that the Rhine plays an enormous role for the UGS and, on the other hand, that the centrality or proximity to university locations as well as the city center is of particular importance for Bonn.

The easy-to-perform example queries show that not only sports types and the respective facilities such as gymnasiums, sports fields or infrastructure such as ping-pong tables can be identified, but also that very precise problems such as littering can be (presumably) identified.

Such and similar queries are particularly important in the context of SG and the smart city, because the information could be forwarded automatically and connected to the respective stakeholders in real time contributing to a local internet of things.

In addition, the enormous importance of seasonal, highly frequented events such as the cherry blossom in Bonn's old town can be seen not only from the terms and hashtags but even from the emojis. Furthermore, the effectiveness of municipal and private advertising campaigns such as the Beethoven Year can also be assessed on the basis of LBSM interaction.

Overall, the possibilities here are almost limitless, not only in thematic but also in spatial terms, and can be applied to numerous contexts which need further investigation.

6.3 Theoretical Findings

Governance in itself is a very vague framework that needs to be concretized as much as possible for individual use cases. The previous definition and conceptualization attempts have shown that there are some overlaps of the concepts GG, SG and MG, but in their combination as MSGG have not yet been addressed in more detail in research. Further, depending on the purpose, it needs to be clarified, whether MSGG should be used as an "analytical, a descriptive and a normative perspective" (HOLTKAMP 2007: 366).

Due to its high abstraction level, it does not serve as an analytical approach. MSGG might be suitable as a normative framework against which not only the dashboard, but also other concrete municipal measures to increase well-being and quality of life can be evaluated. This can only be evaluated in a subsequent field study and speaking with TAYLOR (2016: 4) in the course of implementation by evaluating the respective outcomes. However, in order to be able to assess the extent to which the dashboard has been effective in the course of its use and has been able to deliver a solid information base, more concrete criteria are needed.
The concepts of spatial justice, equity, and equality are an appropriate legitimation for the scope of this thesis. They are aimed at public, equitable resource allocation and need to be explored further through additional studies in order to be evaluated in the course of implementation.

The custom privacy approach (ch. 2.4), consisting of the data instance models of XU et al. (2014: 1151), and the four privacy dimensions of BARKER (2009: 44) in combination with the four post dimensions of LÖCHNER, DUNKEL & BURGHARDT (2018: 2) prove to be a good framework, to evaluate the system immanent privacy. It shows, which instances are to consider, how the amount of data can be reduced efficiently for different purposes and considers the different vulnerability axes of visibility, purpose, granularity and retention which form the crucial information base for a public ethical discussion. It serves well, to get an understanding of what data are actually available to dashboard users. However, it is not suited to evaluate the dashboard infrastructure as a whole system including the computer security approaches (cryptography, access restriction etc.) and what would be available to a hypothetical attacker in the worst-case scenario.

6.4 Future Research

The dashboard developed here is accompanied by the claim that the potential of the massive concentration of data in the hands of LBSM monopolies should not be exploited for commercial purposes or to maximize profits, but rather responsibly used for a socio-spatially equitable distribution of resources in the sense of MSGG. For this purpose, future research should focus primarily on four areas:

1) **Privacy**. HLL and LBSN-Structure provide a very good level of user privacy but cannot yet preserve privacy completely. The additional proposed methods already provide sufficient protection for secure productive use. However, the privacy approach needs to be further developed to eventually meet the DP criteria.
2) **LBSN Dashboard**. The actual dashboard is available as an open-source repository and will be further developed. This thesis will be followed by a pilot study in a municipal setting for which additional visualization features will be developed.
3) **Dashboard Queries and Plugins.** For HLL-DB, aggregate queries from atomic components need further development, particularly for the temporal and social facets. Since the dashboard offers a wide range of analysis possibilities, it is important to develop presets combining terms, hashtags

and emojis etc. that can be easily applied to certain topics such as sports in order to simplify the use as much as possible, but at the same time leave room for highly individualized queries. Further, this thesis intends to encourage the development of open-source plugins based on open data.

4) **Result Interpretation and Evaluation**. The actual meaning of the results must be understood with the analysis of possible biases. The final interpretation of the results is a fundamental step towards successful and democratic policy-making and requires further research.

Ultimately, this thesis wants to be understood as plea for a democratic, ethically responsible and privacy-aware use of LBSM data for smarter municipalities and hopes to have laid a solid foundation for further research.

As declared in the beginning, the dashboard will be presented at three different conferences (see DEUTSCHER VERBAND FÜR ANGEWANDTE GEOGRAPHIE E.V. 2021; LEIBNIZ-INSTITUT FÜR ÖKOLOGISCHE RAUMENTWICKLUNG 2021; VGISCIENCE 2021). Further development of the dashboard as well as the municipal pilot study will be announced in the respective GitHub repository (see WECKMÜLLER 2021b) as well as on the author's personal blog (see WECKMÜLLER 2021c).

References

Literature

ABBASI, A., RASHIDI, T. H., MAGHREBI, M., & WALLER, S. T. (2015). Utilising location based social media in travel survey methods: bringing Twitter data into the play. In A. Pozdnoukhov, D. Sacharidis, & S. Xu (Eds.) Proceedings of the 8th ACM SIGSPATIAL international workshop on location-based social networks (pp. 1-9). The Association for Computing Machinery.

AGOSTINO, D. (2013). Using social media to engage citizens: A study of Italian municipalities. Public Relations Review, 39(3), 232-234.

ALLMENDINGER, P. (2016). Neoliberal spatial governance. Routledge.

ALLMENDINGER, P., & HAUGHTON, G. (2013). The evolution and trajectories of English spatial governance: 'Neoliberal' episodes in planning. Planning Practice & Research, 28(1), 6-26.

ANDERSON, J., CASAS SAEZ, G., ANDERSON, K., PALEN, L., & MORSS, R. (2019). Incorporating context and location into social media analysis: A scalable, cloud-based approach for more powerful data science. In T. X. Bui (Ed.) Proceedings of the 52nd Hawaii International Conference on System Sciences (pp. 2274- 2283). Hawaii International Conference on System Sciences.

ANDREESCU, T., & FENG, Z. (2004). Inclusion-exclusion principle. In T. Andreescu, & Z. Feng (Eds.) A path to combinatorics for undergraduates (pp. 117-141). Birkhäuser.

ANTTIROIKO, A. V. (2015). City branding as a response to global intercity competition. Growth and change, 46(2), 233-252.

ASADI, N., & LIN, J. (2013). Fast candidate generation for real-time tweet search with bloom filter chains. ACM Transactions on Information Systems (TOIS), 31(3), 1-36.

BAIK, J. S. (2020). Data privacy against innovation or against discrimination?: The case of the California Consumer Privacy Act (CCPA). Telematics and Informatics, 52, 1-10.

BARKER, K., ASKARI, M., BANERJEE, M., GHAZINOUR, K., MACKAS, B., MAJEDI, M., PUN, S. & WILLIAMS, A. (2009). A data privacy taxonomy. In A. P. Sexton (Ed.) Dataspace: The Final Frontier – 26th British National Conference on Databases, BNCOD 26 (pp. 42-54). Springer.

BARNS, S. (2018). Smart cities and urban data platforms: Designing interfaces for smart governance. City, culture and society, 12, 5-12.

BAROCAS, S., & NISSENBAUM, H (2014). Big Data's End Run around Anonymity and Consent. In J. Lane, V. Stodden, S. Bender, & H. Nissenbaum (pp. 44-75). Cambridge University Press.

BAUR, N., & BLASIUS, J. (2019). Methoden der empirischen Sozialforschung – Ein Überblick. In N. Baur, & J. Blasius (Eds.) Handbuch Methoden der empirischen Sozialforschung (pp. 1-28). Springer VS.

BERNARD, H. R. (1996). Qualitative data, quantitative analysis. CAM Journal, 8(1), 9-11.

BERNARD, H. R. (2013). Social research methods: Qualitative and quantitative approaches. Sage.

BEVIR, M. (2012). Governance: A very short introduction. OUP Oxford.

BOY, J. D., & UITERMARK, J. (2015). Capture and share the city: Mapping Instagram's uneven geography in Amsterdam. In The Ideal City: between myth and reality – Representations, policies, contradictions and challenges for tomorrow's urban life Urbino (Vol. 27) (pp. 1-20). Research Committee 21 (RC21).

BOYD, D., & CRAWFORD, K. (2012). CRITICAL QUESTIONS FOR BIG DATA. Information, Communication & Society, 15(5), 662–679.

BRONFENBRENNER, M. (1973). Equality and equity. The ANNALS of the American Academy of Political and Social Science, 409(1), 9-23.

BUCKLEY, F. H. (2009). Fair governance: paternalism and perfectionism. Oxford University Press.

BURTON, S. H., TANNER, K. W., GIRAUD-CARRIER, C. G., WEST, J. H., & BARNES, M. D. (2012). "Right Time, Right Place" Health Communication on Twitter: Value and Accuracy of Location Information. Journal of medical Internet research, 14(6), e156.

CAO, G., WANG, S., HWANG, M., PADMANABHAN, A., ZHANG, Z., & SOLTANI, K. (2015). A scalable framework for spatiotemporal analysis of location-based social media data. Computers, Environment and Urban Systems, 51, 70-82.

CHEFFINS, B. R. (2013). The history of corporate governance. The Oxford handbook of corporate governance, 46, 56-58.

CHEN, L., HUANG, Y., OUYANG, S., & XIONG, W. (2021). The Data Privacy Paradox and Digital Demand. National Bureau of Economic Research Working Paper, 28854, 1-53.

DAHLGAARD, S., KNUDSEN, M. B. T., & THORUP, M. (2017). Practical hash functions for similarity estimation and dimensionality reduction. arXiv preprint. https://arxiv.org/pdf/1711.08797.pdf, 1-19.

DAILEY-O'CAIN, J. (2017). Trans-national English in social media communities. Springer.

DALENIUS, T. (1977). Towards a methodology for statistical disclosure control. Statistik Tidskrift, 15(5), 429-444.

DALVI, N., OLTEANU, M., RAGHAVAN, M., & BOHANNON, P. (2014). Deduplicating a places database. In Proceedings of the 23rd international conference on World wide web (pp. 409-418). Association for Computing Machinery.

DAMGÅRD, I. B. (1989). A design principle for hash functions. In G. Brassard (Ed.) Advances in Cryptology - CRYPT0 '89, LNCS 435 (416-427). (New York, NY) Springer.

DAMKOWSKI, W., & RÖSENER, A. (2004). Good Governance auf der lokalen Ebene. Arbeitspapiere für Staatswissenschaft, 9, n.pp..

DASGUPTA, A., LANG, K., RHODES, L., & THALER, J. (2015). A framework for estimating stream expression cardinalities. arXiv preprint. https://arxiv.org/pdf/1510.01455.pdf, 1-33.

DE CAPITANI DI VIMERCATI, S., FORESTI, S., LIVRAGA, G., & SAMARATI, P. (2012). Data privacy: definitions and techniques. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 20(06), 793-817.

DE MAURO, A., GRECO, M., & GRIMALDI, M. (2015). What is big data? A consensual definition and a review of key research topics. In G. Giannakopoulos, D. P. Sakas, & D. Kyriaki-Manessi (Eds.) AIP Conference Proceedings 1644(1), pp. 97-104). American Institute of Physics.

DEACON, B., OLLILA, E., KOIVUSALO, M., & STUBBS, P. (2003). Global social governance – Themes and Prospects. Ministry for Foreign Affairs of Finland.

DESFONTAINES, D., LOCHBIHLER, A. & BASIN, D. (2019). Cardinality Estimators do not Preserve Privacy. Proceedings on Privacy Enhancing Technologies, 2019(2), 26-46.

DI MININ, E., FINK, C., HAUSMANN, A., KREMER, J., & KULKARNI, R. (2021). How to address data privacy concerns when using social media data in conservation science. Conservation Biology, 35(2), 437-446.

DOLLERY, B., & JOHNSON, A. (2005). Enhancing efficiency in Australian local government: An evaluation of alternative models of municipal governance. Urban Policy and Research, 23(1), 73-85.

DÖRING, J., & THIELMANN, T. (2015). Spatial turn: das Raumparadigma in den Kultur-und Sozialwissenschaften. transcript Verlag.

DUNKEL, A. (2016). Assessing the perceived environment through crowdsourced spatial photo content for application to the fields of landscape and urban planning. [Doctoral dissertation, TU Dresden]. CORE open access research.

DUNKEL, A., ANDRIENKO, G., ANDRIENKO, N., BURGHARDT, D., HAUTHAL, E., & PURVES, R. (2019). A conceptual framework for studying collective reactions to events in location-based social media. International Journal of Geographical Information Science, 33(4), 780-804.

DUNKEL, A., LÖCHNER, M., & BURGHARDT, D. (2020). Privacy-Aware Visualization of Volunteered Geographic Information (VGI) to Analyze Spatial Activity: A Benchmark Implementation. ISPRS International Journal of Geo-Information, 9(10), 1-21.

DWORK, C. (2008). Differential privacy: A survey of results. In M. Agrawal, D.-Z. Du, Z. Duan, & A. Li, (Eds.) International conference on theory and applications of models of computation (pp. 1-19). Springer.

DWORKIN, R. (1981). What is equality? Part 1: Equality of welfare. Philosophy & public affairs, 10(3), 185-246.

EISENSTEIN, J. (2019). Introduction to natural language processing. MIT press.

ERTL, O. (2017). New cardinality estimation algorithms for hyperloglog sketches. arXiv preprint. https://arxiv.org/pdf/1702.01284.pdf, 1-56.

EUBANKS, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

EVANS, L., & SAKER, M. (2017). Location-based social media: Space, time and identity. Springer.

FEYISETAN, O., DRAKE, T., BALLE, B., & DIETHE, T. (2019). Privacy-preserving active learning on sensitive data for user intent classification. arXiv preprint. https://arxiv.org/pdf/1903.11112.pdf, 1-9.

FIALLOS, A., JIMENES, K., FIALLOS, C., & FIGUEROA, S. (2018). Detecting topics and locations on Instagram photos. In L. Terán, & A. Meier (Eds.) 2018 International Conference on eDemocracy & eGovernment (ICEDEG) (pp. 246-250). The Institute of Electrical and Electronics Engineers.

FISCHER, F. (2008). Location Based Social Media – Considering the Impact of Sharing Geographic Information on Individual Spatial Experience. In A. Car, G. Griesebner, & J. Strobl (Eds.) Geospatial Crossroads @ GI_Forum '08. Proceedings of the Geoinformatics Forum Salzburg (pp. 1-7). Wichmann.

FISHER, D. (2004). National governance and the global climate change regime. Rowman & Littlefield.

FLAJOLET, P., & MARTIN, G. N. (1985). Probabilistic counting algorithms for data base applications. Journal of computer and system sciences, 31(2), 182-209.

FLAJOLET, P., FUSY, É., GANDOUET, O., & MEUNIER, F. (2007). Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. Analysis of Algorithms 2007 (AofA07), 127–146.

FOLGER, R., SHEPPARD, B. H., & BUTTRAM, R. T. (1995). Equity, equality, and need: Three faces of social justice. In B. B. Bunker & J. Z. Rubin (Eds.) The Jossey-Bass management series and The Jossey-Bass conflict resolution series. Conflict, cooperation, and justice: Essays inspired by the work of Morton Deutsch (pp. 261–289). Jossey-Bass/Wiley.

FOSTER, K. W. (2006). Improving municipal governance in China: Yantai's pathbreaking experiment in administrative reform. Modern China, 32(2), 221-250.

FRANKENFELD, P. J. (1992). Technological citizenship: A normative framework for risk studies. Science, Technology, & Human Values, 17(4), 459-484.

FÜRST, D. (2004). Regional governance. In A. Benz A. (Ed.) Governance – Regieren in komplexen Regelsystemen. Governance (pp. 45-64). Springer VS.

GAO, H., TANG, J., & LIU, H. (2012). gSCorr: Modeling geo-social correlations for new check-ins on location-based social networks. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 1582-1586). The Association for Computing Machinery.

GARCIA-RECUERO, A. (2021). Approximate Privacy-Preserving Neighbourhood Estimations. arXiv preprint https://arxiv.org/pdf/arXiv:2102.12610.pdf, 1-5.

GEBHARDT, H., & REUBER, P. (2011). Aktuelle Leitlinien der Strukturierung und Entwicklung der Humangeographie. In H. Gebhardt (Ed.) Geographie: Physische Geographie und Humangeographie (Vol. 2) (pp. 645-653). Spektrum Akademischer Verlag.

GEDIK, B., & LIU, L. (2004). A customizable k-anonymity model for protecting location privacy. Georgia Institute of Technology.

GEHRKE, J., LUI, E., & PASS, R. (2011). Towards privacy for social networks: A zero-knowledge based definition of privacy. In Y. Ishai (Ed.) Theory of cryptography conference (pp. 432-449). Springer.

GISSELQUIST, R. M. (2012). Good governance as a concept, and why this matters for development policy. WIDER Working Paper 2012(30), 1-36.

GONG, Z., SUN, G. Z., & XIE, X. (2010). Protecting privacy in location-based services using k-anonymity without cloaked region. In 2010 Eleventh International Conference on Mobile Data Management (pp. 366-371). The Institute of Electrical and Electronics Engineers.

GRAHAM, J., PLUMPTRE, T. W., & AMOS, B. (2003). Principles for good governance in the 21st century. In J. Graham, B. Amos, & T. Plumptre (Eds.) Policy Brief No.15 (pp. 1-6). Institute On Governance.

GRUBER, J. F., & HOLSTEIN, J.A. (2014): Analytic inspiration in ethnographic fieldwork. In: U. Flick (Ed.) The SAGE handbook of qualitative data analysis (pp. 35-48). Sage.

GÜNZEL, S. (2010). Raum: ein interdisziplinäres Handbuch. Metzler.

HAQ, S. M. A. (2011): Urban Green Spaces and an Integrative Approach to Sustainable Environment. Journal of Environmental Protection 2(5). 601–608.

HARVEY, D. (2010). Social justice and the city (Vol. 1). University of Georgia Press.

HASAN, S., ZHAN, X., & UKKUSURI, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In Proceedings of the 2nd ACM SIGKDD international workshop on urban computing (pp. 1-8). Association for Computing Machinery.

HASNAT, M. M., & HASAN, S. (2018). Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. Transportation Research Part C: Emerging Technologies, 96, 38-54.

HAZARI, S., & BROWN, C. (2013). An empirical investigation of privacy awareness and concerns on social networking sites. Journal of Information Privacy and Security, 9(4), 31-51.

HOCHMAN, N., & MANOVICH, L. (2013). Zooming into an Instagram City: Reading the local through social media. First Monday.

HOLLANDS, R. G. (2008). Will the real smart city please stand up?. City – analysis of urban trends, culture, theory, policy, action 12(3), 303-320.

HOLTKAMP, L. (2007). Local governance. In A. Benz., S. Lütz, U. Schimank, & G. Simonis (Eds.) Handbuch Governance (pp. 366-377). Springer VS.

HUANG, B., & CARLEY, K. M. (2019). A large-scale empirical study of geotagging behavior on twitter. In F. Spezzano, W. Chen, & X. Xiao (Eds.) Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 365-373). Institute of Electrical and Electronics Engineers.

ILIEVA, R. T., & MCPHEARSON, T. (2018). Social-media data for urban sustainability. Nature Sustainability, 1(10), 553-565.

ISAAK, J., & HANNA, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. Computer, 51(8), 56-59.

JANGIRALA, S., & CHAKRAVARAM, V. (2021). Authenticated and Privacy Ensured Smart Governance Framework for Smart City Administration. In A. Kumar, & S. Mozar (Eds.) Proceedings of the

3rd International Conference on Communications and Cyber Physical Engineering (pp. 931-942). Springer.

JIANG, B., & MIAO, Y. (2015). The evolution of natural cities from the perspective of location-based social media. The Professional Geographer, 67(2), 295-306.

JOFFE-NOTIER, T. (2020). An Inquiry into Equity and Democracy. In D. Forde, A. Bouque, E. A. Kahn, T. M. McCann, & C. Walter (Eds.) Inquiry Units for English Language Arts: Inspiring Literacy Learning, Grades 6-12 (pp. 109-136). Rowman & Littlefield.

KESSLER, C., & McKenzie, G. (2018). A geoprivacy manifesto. Transactions in GIS, 22(1), 3-19.

KJÆR, A. (2004). Governance. Wiley.

KOUNADI, O., & LEITNER, M. (2014). Why does geoprivacy matter? The scientific publication of confidential data presented on maps. Journal of Empirical Research on Human Research Ethics, 9(4), 34-45.

KUCKARTZ, U. (2014). Mixed Methods: Methodologie, Forschungsdesigns und Analyseverfahren. Springer.

LEPPÄNEN, S., & KYTÖLÄ, S. (2017). Investigating multilingualism and multisemioticity as communicative resources in social media. In M. Martin-Jones, & D. Martin (Eds.) Researching multilingualism: Critical and ethnographic perspectives (pp. 155–171). Routledge.

LI, Z., SHARMA, V., & MOHANTY, S. P. (2020). Preserving data privacy via federated learning: Challenges and solutions. Consumer Electronics Magazine, 9(3), 8-16.

LIM, K. H., CHAN, J., KARUNASEKERA, S., & LECKIE, C. (2019). Tour recommendation and trip planning using location-based social media: A survey. Knowledge and Information Systems, 60(3), 1247-1275.

LIN, J., OENTARYO, R., LIM, E. P., VU, C., VU, A., & KWEE, A. (2016). Where is the goldmine? Finding promising business locations through Facebook data analytics. In Proceedings of the 27th ACM Conference on Hypertext and Social Media (pp. 93-102). Association for Computing Machinery.

LIU, L. (2007). From data privacy to location privacy: models and algorithms. In VLDB '07: Proceedings of the 33rd International Conference on Very Large Data Bases (pp. 1429-1430). VLDB Endowment.

LIU, X., LIU, K., GUO, L., LI, X., & FANG, Y. (2013). A game-theoretic approach for achieving k-anonymity in location based services. In The 32nd IEEE International Conference on Computer Communications (pp. 2985-2993). The Institute of Electrical and Electronics Engineers.

LÖCHNER, M., DUNKEL, A., & BURGHARDT, D. (2018). A privacy-aware model to process data from location-based social media. In D. Burghardt, S. Chen, G. Andrienko, Gennady, N. Andrienko, R. Purves, & A. Diehl (Eds.) VGI Geovisual Analytics Workshop, colocated with BDVA 2018 (pp. 1-5). The Institute of Electrical and Electronics Engineers.

LÖCHNER, M., FATHI, R., SCHMID, D., DUNKEL, A., BURGHARDT, D., FIEDRICH, F., & KOCH, S. (2020). Case study on privacy-aware social media data processing in disaster management. ISPRS International Journal of Geo-Information, 9(12), 1-13.

LUCY, W. (1981). Equity and planning for local services. Journal of the American Planning Association, 47(4), 447-457.

MÄDING, H. (2021). Gleichwertige Lebensverhältnisse und Fachpolitik – explorative Beobachtungen und Überlegungen am Beispiel der aktuellen Kohlepolitik. Raumforschung und Raumordnung| Spatial Research and Planning, 79(1), 73-86.

MCKENZIE, G., JANOWICZ, K., & SEIDL, D. (2016). Geo-privacy beyond coordinates. In T. Sarjakoski, M. J. Santos, L. T. Sarjakoski (Eds.) Geospatial Data in a Changing World (pp. 157-175). Springer.

METCALF, J., & CRAWFORD, K. (2016). Where are human subjects in Big Data research? The emerging ethics divide. Big Data & Society, 3(1), 1-14.

MOORE, A. D. (2008). Defining privacy. Journal of Social Philosophy, 39(3), 411-428.

MORSTATTER, F., PFEFFER, J., LIU, H., & CARLEY, K. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 7, No. 1) (pp. 400-408). The AAAI Press.

NANDA, V. P. (2006). The "Good Governance" Concept Revisited. The ANNALS of the American Academy of Political and Social Science, 603(1), 269–283.

NIU, B., LI, Q., ZHU, X., CAO, G., & LI, H. (2014). Achieving k-anonymity in privacy-aware location-based services. In The 33rd Annual IEEE International Conference on Computer Communications (INFOCOM'14) (pp. 754-762). The Institute of Electrical and Electronics Engineers.

PADMANABHAN, A., WANG, S., CAO, G., HWANG, M., ZHANG, Z., GAO, Y., SOLTANI, K. & LIU, Y. (2014). FluMapper: A cyberGIS application for interactive analysis of massive location-based social media. Concurrency and computation: Practice and Experience, 26(13), 2253-2265.

PELLERT, M., LASSER, J., METZLER, H., & GARCIA, D. (2020). Dashboard of sentiment in Austrian social media during COVID-19. arXiv preprint. https://arxiv.org/pdf/2006.11158.pdf, 1-23.

PEREIRA, G. V., PARYCEK, P., FALCO, E., & KLEINHANS, R. (2018). Smart governance in the context of smart cities: A literature review. Information Polity, 23(2), 143-162.

POLITOU, E., ALEPIS, E., & PATSAKIS, C. (2018). Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions. Journal of Cybersecurity, 4(1), 1–20.

RANTANEN, H., & KAHILA, M. (2009). The SoftGIS approach to local knowledge. Journal of environmental management, 90(6), 1981-1990.

REUTER, C., LUDWIG, T., KAUFHOLD, M. A., & PIPEK, V. (2015). XHELP: Design of a cross-platform social-media application to support volunteer moderators in disasters. In CHI '15: CHI Conference on Human Factors in Computing Systems (pp. 4093-4102). Association for Computing Machinery.

REVIRIEGO, P., & TING, D. (2020). Security of HyperLogLog (HLL) cardinality estimation: Vulnerabilities and protection. The Institute of Electrical and Electronics Engineers Communications Letters, 24(5), 976-980.

REYMAN, J. (2013). User data on the social web: Authorship, agency, and appropriation. College English, 75(5), 513-533.

RHODES, R. A. (2007). Understanding governance: Ten years on. Organization studies, 28(8), 1243-1264.

ROSE, R., & PEIFFER, C. (2018). Bad governance and corruption. Springer.

SAMARATI, P., & SWEENEY, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. [Technical Report] SRI-CSL-98-04, SRI Computer Science Laboratory, 1-19.

SANTISO, C. (2001). Good governance and aid effectiveness: The World Bank and conditionality. The Georgetown public policy review, 7(1), 1-22.

SCHÖLER, G., & WALTHER, C. (2003). A Practical Guidebook on Strategic Management for Municipal Administration. The World Bank, Bertelsmann Foundation.

SCHOLL, H. J., & ALAWADHI, S. (2016). Creating Smart Governance: The key to radical ICT overhaul at the City of Munich. Information Polity, 21(1), 21-42.

SCHOLL, H. J., & SCHOLL, M. C. (2014). Smart Governance: A Roadmap for Research and Practice. In M. Kindling, & E. Greifeneder (Eds.) The iConference2014 Proceedings (p. 163–176). iSchools.

SCHREURS, L., & VANDENBOSCH, L. (2020). Introducing the Social Media Literacy (SMILE) model with the case of the positivity bias on social media. Journal of Children and Media, [no vol., no iss.], 1-18.

SCHWARTZ, R., & HALEGOUA, G. R. (2015). The spatial self: Location-based identity performance on social media. New media & society, 17(10), 1643-1660.

SEN, A. (1980). Equality of what?. In S. McMurrin (Ed.) The Tanner Lectures on 3021 Human Values, Salt Lake City (pp. 197-220). University of Utah Press.

SHANG, S., GUO, D., LIU, J., ZHENG, K., & WEN, J. R. (2016). Finding regions of interest using location based social media. Neurocomputing, 173, 118-123.

SMEDBY, N., & QUITZAU, M. B. (2016). Municipal governance and sustainability: The role of local governments in promoting transitions. Environmental Policy and Governance, 26(5), 323-336.

SMITH, B. C. (2007). Good governance and development. Macmillan International Higher Education.

SOJA, E. W. (2013). Seeking spatial justice (Vol. 16). University of Minnesota Press.

SOJA, E. W. (2015). Vom „Zeitgeist“ zum „Raumgeist “ – New Twists on the Spatial Turn. In J. Döring, & T. Thielmann (Eds.) Spatial turn: das Raumparadigma in den Kultur-und Sozialwissenschaften (pp. 241-262). transcript Verlag.

STRÜVER, A. (2011). Der kleine Unterschied und seine großen Folgen – Humangeographische Perspektiven durch die Kategorie Geschlecht. In H. Gebhardt (Ed.) Geographie: Physische Geographie und Humangeographie (Vol. 2) (pp. 667-675). Spektrum Akademischer Verlag.

SWEENEY, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

TANG, W., REN, J., & ZHANG, Y. (2018). Enabling trusted and privacy-preserving healthcare services in social media health networks. IEEE Transactions on Multimedia, 21(3), 579-590.

TAYLOR, Z. T. (2016). Good governance at the local level: Meaning and measurement. In P. Campsie, & S. Zhang (Eds.) IMFG Papers on Municipal Finance and Governance (pp. 1-39). Institute on Municipal Finance & Governance.

TENKANEN, H., DI MININ, E., HEIKINHEIMO, V., HAUSMANN, A., HERBST, M., KAJALA, L., & TOIVONEN, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. Scientific reports, 7(1), 1-11.

TSOU, M. H., JUNG, C. T., ALLEN, C., YANG, J. A., GAWRON, J. M., SPITZBERG, B. H., & HAN, S. (2015). Social media analytics and research test-bed (SMART dashboard). In A. Gruzd, J. Jacobson, P.

Mai, & B. Wellman (Eds.) Proceedings of the 2015 international conference on social media & society (pp. 1-7). Association for Computing Machinery.

TSOU, M. H. & LEITNER, M. (2013) Visualization of social media: seeing a mirage or a message?, Cartography and Geographic Information Science, 40(2), 55-60.

UNDP (1997): Governance for Sustainable Development. Community Organization, Training, Research, and Advocacy Institute (CO-TRAIN), & United Nations Development Programme (UNDP).

VALKENBURG, G. (2012). Sustainable technological citizenship. European Journal of Social Theory, 15(4), 471–487.

WANG, J., CAI, Z., LI, Y., YANG, D., LI, J., & GAO, H. (2018). Protecting query privacy with differentially private k-anonymity in location-based services. Personal and Ubiquitous Computing, 22(3), 453-469.

WANG, S., & SINNOTT, R. O. (2017). Protecting personal trajectories of social media users through differential privacy. Computers & Security, 67, 142-163.

WERLEN, B. (2008). Sozialgeographie: eine Einführung (Vol 3). UTB.

WILKEN, R. (2014). Places nearby: Facebook as a location-based social media platform. New Media & Society, 16(7), 1087-1103.

WILLKE, H. (2006). Global governance. transcript Verlag.

WILLKE, H. (2007). Smart governance: governing the global knowledge society. Campus Verlag.

WOLCH, J. R., BYRNE, J., & J. P. NEWELL (2014): Urban green space, public health, and environmental justice: The challenge of making cities “just green enough“. Landscape and Urban Planning 125, 234–244.

WORLD BANK. (1991). Managing Development: The Governance Dimension, a Discussion Paper. World Bank.

WORLD BANK. (1992). Governance and development. The World Bank.

XU, L., JIANG, C., WANG, J., YUAN, J., & REN, Y. (2014). Information security in big data: privacy and data mining. IEEE Access, 2, 1149-1176.

YI, Y., LIU, Y., HE, H., & LI, Y. (2012). Environment, governance, controls, and radical innovation during institutional transitions. Asia Pacific Journal of Management, 29(3), 689-708.

YOUDE, J. (2012). Global health governance. Polity.

YUAN, Y., WEI, G., & LU, Y. (2018). Evaluating gender representativeness of location-based social media: A case study of Weibo. Annals of GIS, 24(3), 163-176.

ZHAO, P., LI, J., ZENG, F., XIAO, F., WANG, C., & JIANG, H. (2018). ILLIA: Enabling k-anonymity-based privacy preserving against location injection attacks in continuous LBS queries. IEEE Internet of Things Journal, 5(2), 1033-1042.

List of Web References

All web references were last successfully accessed on June 21, 2021, 8:00 to 9:00 AM.

AGAFONKIN, V. & CONTRIBUTORS (2021). Leaflet. https://github.com/Leaflet/Leaflet.

APPLEBY, A. (2021). MurmurHash3. https://github.com/aappleby/smhasher/blob/master/src/MurmurHash3.cpp.

BEETHOVEN JUBILÄUMS GMBH (2021). Über BTHVN2020. https://www.bthvn2020.de/ueber-uns/ueber-bthvn2020.

BUNDESSTADT BONN (2021a). Smart City Bonn. https://www.bonn.de/microsite/smartcity/.

BUNDESSTADT BONN (2021b). Fläche Stadtgebiet Bonn. https://opendata.bonn.de/dataset/fl%C3%A4che-stadtgebiet-bonn.

BUNDESSTADT BONN (2021c). Region Bonn: Tourismus-Jahresbilanz 2019 mit neuem Rekord. https://www.bonn.de/pressemitteilungen/februar/region-bonn-tourismus-bilanz-2019-mit-neuem-rekord.php.

BUNDESSTADT BONN (2021d). Offene Daten Bonn. https://opendata.bonn.de/.

BUNDESSTADT BONN STATISTIKSTELLE (2021). Statistik aktuell Februar 2021 – Bevölkerung in der Bundesstadt Bonn Stichtag 31.12.2020. https://www2.bonn.de/statistik/dl/ews/Bevoelkerungsstatistik2020.pdf.

BURK, A., HUHN, L., & WECKMÜLLER, D. (2020). Applying an adjusted Cultural Ecosystem Services concept on spatiotemporal perception and non-material usage patterns of Urban Green Spaces – A social media data analysis on the example of Bonn. https://github.com/do-me/instagreens-bonn/raw/master/Instagreens-Bonn.pdf.

CITUS & CONTRIBUTORS (2021). Postgresql-hll. https://github.com/citusdata/postgresql-hll.

DEUTSCHER VERBAND FÜR ANGEWANDTE GEOGRAPHIE E.V. (2021). #GeoWoche2021 / Deutscher Kongress für Geographie (DKG). https://geographie-dvag.de/veranstaltungen/dkg/.

DUNKEL, A. (2020). HyperLogLog for Real Time HyperLogLog for Real Time LBSM Visual Analytics – A solution to the count distinct problem and privacy? https://ad.vgiscience.org/lbsn-hll-slides.

DUNKEL, A., LÖCHNER, M., KRUMPE, F. & Contributors (2021). LBSN Structure. https://lbsn.vgiscience.org/.

DUNKEL, A. & LÖCHNER M. (2021a). LBSN HLL Database - Docker Container. https://gitlab.vgiscience.de/lbsn/databases/hlldb/.

DUNKEL, A. & LÖCHNER M. (2021b). Lbsntransform. https://lbsn.vgiscience.org/lbsntransform/docs/.

DUNKEL, A. & LÖCHNER M. (2021c). Importing lbsntransform as a package. https://lbsn.vgiscience.org/lbsntransform/docs/package/#importing-lbsntransform-as-a-package.

DUNKEL, A. & LÖCHNER M. (2021d). Input Mappings. https://lbsn.vgiscience.org/lbsntransform/docs/mappings/

FACEBOOK (2021a). Facebook for Business – Tell your brand story your way with Instagram. https://www.facebook.com/business/marketing/instagram.

FACEBOOK (2021b). Facebook Places. https://www.facebook.com/places/.

FLICKR (2009). 100,000,000 geotagged photos (plus). https://code.flickr.net/2009/02/04/100000000-geotagged-photos-plus/.

FLICKR (2015). Find every photo with Flickr's new unified search experience. https://blog.flickr.net/en/2015/05/07/flickr-unified-search/.

FLICKR (2021a). The Flickr Developer Guide. https://www.flickr.com/services/developer/.

FLICKR (2021b). Work at Flickr. https://www.flickr.com/jobs/.

FLICKR (2021c). Flickr map. https://www.flickr.com/map.

FLICKR (2021d). The App Garden. https://www.flickr.com/services/api/explore/flickr.photos.search.

GEOGRAPHISCHES INSTITUT UNIVERSITÄT BONN (2020). Was passiert eigentlich in der vorlesungsfreien Zeit? - Teil 16. https://www.facebook.com/GeographischesInstitutUniBonn/posts/was-passiert-eigentlich-in-der-vorlesungsfreien-zeit-teil-16studieren-im-hofgart/1703976159767795/.

GIS CERTIFCATION INSTITUTE (2003). A GIS Code of Ethics. https://www.gisci.org/Ethics/CodeofEthics.aspx.

INSTAGRAM (2021a). Terms of Use, revised: December 20, 2020. https://help.instagram.com/581066165581870.

INSTAGRAM (2021b). Instagram locations. https://www.instagram.com/explore/locations/.

LEIBNIZ-INSTITUT FÜR ÖKOLOGISCHE RAUMENTWICKLUNG (2021). DFNS 2021 – Dresdner Flächennutzungssymposium. http://dfns2021.ioer.info/.

LIN, J., OENTARYO, R., LIM, E. P., VU, C., VU, A., & KWEE, A. (2021). Business Analytics. https://research.larc.smu.edu.sg/bizanalytics/.

MAYNOOTH UNIVERSITY (2021). Dublin Dashboard. https://www.dublindashboard.ie/.

OPENSTREETMAP WIKI (2021). SymbolsTab. https://wiki.openstreetmap.org/wiki/SymbolsTab.

PEW RESEARCH CENTER (2021). Social Media Use in 2021. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/.

RAMÍREZ, S. & CONTRIBUTORS (2021). FastApi. https://github.com/tiangolo/fastapi.

ROBINSON, R. (2016). Why Smart Cities still aren't working for us after 20 years. And how we can fix them. https://theurbantechnologist.com/2016/02/01/why-smart-cities-still-arent-working-for-us-after-20-years-and-how-we-can-fix-them/.

SENDIBLE (2021). Instagram Locations: Why Adding Them is Always a Good Idea. https://www.sendible.com/insights/instagram-locations.

TWITTER (2021a). Letter to Shareholders - Q4 and Fiscal Year 2020. https://s22.q4cdn.com/826641620/files/doc_financials/2020/q4/FINAL-Q4'20-TWTR-Shareholder-Letter.pdf.

TWITTER (2021b). How to add your location to a Tweet. https://help.twitter.com/en/using-twitter/tweet-location.

TWITTER (2021c) Take your research further with Twitter data. https://developer.twitter.com/en/solutions/academic-research.

VGISCIENCE (2021). VGIscience 2021 lecture series. https://www.vgiscience.org/lecture-series.html

WECKMÜLLER, D. (2021a). LBSN-Dashboard. https://github.com/do-me/LBSN-Dashboard.

WECKMÜLLER, D. (2021b). LBSN-Thesis. https://github.com/do-me/LBSN-Thesis.

WECKMÜLLER, D. (2021c) geo.rocks. https://geo.rocks.

WECKMÜLLER, D. (2021d). Fast-Instagram-Scraper. https://github.com/do-me/fast-instagram-scraper.

WECKMÜLLER, D. (2021e). Using HyperLogLog with Leaflet-Hexbins. https://geo.rocks/post/hexbins-js-hll/.

List of Abbreviations

Abbreviation	Meaning
AI	artificial intelligence
AOI	area of interest
API	application programming interface
art.	article
ch.	chapter
CPU	central processing unit
CSV	comma separated values
CSS	cscading style sheets
DAU	daily active users
DP	differential privacy
eq.	equation
fig.	figure
GDPR	General Data Protection Regulation
GG	good governance
GUI	graphical user interface
hexbins	hexagonal bins
HLL	HyperLogLog
HLL-DB	HyperLogLog database
ID	Identifier
JSON	JavaScript Object Notation
LBSM	location-based social media
LBSN	location-based social network
LSLZ	longest series of leading zeros
LZ	leading zeros
MG	municipal governance
ML	machine learning
MSGG	municipal smart good governance
n.p./pp.	no page/s
NLP	natural language processing
o.s.	or similar
OSM	OpenStreetMap
POI	point of interest
SG	smart governance
SM	social media

SN	social network
tab.	table
UGS	urban green spaces
UID	unique identifier
UN	United Nations
UNDP	United Nations Development Program
VGI	volunteered geographic information

List of Symbols

Abbreviation	Meaning	Dimension
e	binary sample event series	-
m	number of HLL registers	-
$N(e)_{est}$	estimated number of runs	-
$N(e)_{real}$	real number of runs	-
$P(Z(e))$	leading zeros probability	%
$Z(e)$	number of leading zeros of e	-

List of Figures

List of Tables

List of Equations

List of Regulations