Capturing Vernacular Geography from Georeferenced Tags

Master Thesis

Livia Hollenstein

Faculty Representative:

Prof. Dr. Robert Weibel

Supervisors:

Dr. Ross Purves
Dr. Clare Davies

Institute of Geography
University of Zurich

November 2008

Persönliche Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und die den verwendeten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Zürich, 28. November 2008

Livia Hollenstein

Acknowledgments

I would like to thank my supervisor, Dr. Ross Purves, for his help and encouragement during the course of my work on this study. Furthermore, I would also like to thank Dr. Clare Davies for her valuable advice and Prof. Dr. Robert Weibel for supporting this Master project.

Finally, I wish to thank the following people who offered advice and perspective at various stages of this project: Markus, Othmar, Flurina, Janine, Daria as well as my parents for their support and words of encouragement.

Abstract

In view of the abundance of geographically related information on the world wide web and the ubiquity of location-aware devices, today the majority of seekers as well as providers of spatially related information are not experts in the geographic domain. Humans acquire extensive spatial knowledge in the course of life, but there is a lot of vagueness inherent in the way we conceive and refer to geographic location. Instead of exact distances and coordinates, we employ vague spatial concepts such as 'downtown' or 'Soho' without being concerned about exact boundaries. Hence, in GIScience attention was recently drawn to the automated interpretation of arbitrarily employed place names, which are challenging and difficult to interpret and process by computer systems.

This study explores the abundance of absolute references between place and associated descriptions available from georeferenced items in online tagging systems as a source of knowledge about vernacular geography. It draws on voluntarily created keywords to categorise and describe georeferenced photos in the online photo-sharing platform Flickr, a process which has been considered as a proxy of how people intuitively refer to location. Flickr, featuring about ninety million georeferenced photos at the time of writing, is characteristic for the highly user-centric structure of the web, which provides locative aspects beyond the digital map data traditionally used in GIS. It is an issue of intensive research as to extract, mine, visualise, and exploit this geographically relevant information.

In view of the availability of a global and multilingual database, the investigation of spatial terminology applied in terms of tags covered German as well as English language areas. It focussed on generic concepts employed to describe the urban core, which is considered as a prototype of a vague geographic entity. Different visualisation techniques as well as a standard GIS method, including an automatic way to deal with potential outliers, were adopted in order to investigate the bias introduced by different user groups, to establish global and local patterns of place tag-usage, and to derive vague footprints from georeferenced place semantics.

The quantitative evaluation of tag-usage revealed that a large portion of the top-ranked tags in georeferenced Flickr samples correspond to place names. The city name was identified as the granularity level people most intuitively think of when assigning locational information. The fraction of generic city core terms is generally marginal and the majority of such tags tend to exhibit problematic values with regard to user ubiquity, while particularly in larger cities, specific place names of areas and neighbourhoods are common. The nature of derived vague footprints suggest that the average user has a distinct idea of specific places and that the users' attitude towards the creation of metadata meet the requirements for practical purposes at the sub-city level of granularity. Despite the availability of an immense amount of empirical data, the analysis reveals a considerable bias in terms of user contribution as well as in spatial distribution, a fact which must be accounted for carefully during evaluation. Given a 'critical mass' of items and participating users, the results confirm that the abundance and quality of formal and textual place references in Flickr hold relevant information for the geographic discipline and are highly interesting for the extraction of common-sense spatial knowledge.

Zusammenfassung

Angesichts der Fülle geographisch relevanter Information im Internet und der zunehmenden Verbreitung ortungsfähiger, mobiler Technologien sind die Mehrheit der Informationssuchenden und -anbietern heutzutage nicht Experten der geographischen Disziplin. Menschen eignen sich im Laufe des Lebens beachtliches räumliches Wissen an; trotzdem ist die Art, wie wir den geographischen Raum konzeptualisieren und beschreiben typischerweise von Vagheit geprägt. Statt exakter Distanzen und Koordinaten benutzen wir unklar definierte Konzepte wie 'Innenstadt' oder 'Soho', ohne uns um den Verlauf von exakten Grenzen zu kümmern. Daher wurde in GIScience die Aufmerksamkeit vermehrt auf die automatische Interpretation von beliebig verwendeten Ortsnamen gezogen, welche von Computersystemen nur schwer interpretier- und verarbeitbar sind.

Die vorliegende Arbeit untersucht direkte Referenzen zwischen Orten und assoziierten Beschreibungen in web-basierten Tagging-Systemen als Informationsquelle über umgangssprachliche Geographie. Sie stützt sich auf Schlüsselwörter, so genannte Tags, die von Benutzern freiwillig erzeugt werden, um den Inhalt georeferenzierter Bilder in der photosharing Plattform Flickr zu kategorisieren und zu beschreiben. Dieser Prozess wird als Modell dafür betrachtet, wie Menschen Orte intuitiv beschreiben. Die zurzeit etwa 90 Millionen georeferenzierten Photos und ihre Metadaten auf Flickr sind charakteristisch für das benutzerzentrierte Internet, welches räumliche Aspekte beinhaltet, die über die traditionellerweise in GIS benutzten Kartendaten hinausgehen. Die Extraktion, Visualisierung und Nutzung dieser geographisch relevanten Daten ist Gegenstand der aktuellen Forschung.

Angesichts der Verfügbarkeit einer weltweiten und mehrsprachigen Datengrundlage deckt die Untersuchung räumlicher Terminologie, wie sie in Form von Tags verwendet wird, sowohl den deutschen als auch den englischen Sprachraum ab. Die Analyse konzentriert sich auf generische Konzepte zur Bezeichnung des Stadtzentrums, welches als Prototyp einer vagen und unklar definierten räumlichen Einheit betrachtet wird. Verschiedene Visualisierungstechniken und eine standardisierte GIS-Methode werden verwendet, um die Verzerrung der Daten durch einzelne Benutzer, sowie weltweite und lokale Muster von Orts-Tags als auch die aus georeferenzierter Ortsbeschreibung abgeleiteten Repräsentationen, so genannte 'Footprints', zu analysieren.

Eine quantitative Analyse der Verwendung von Tags hat gezeigt, dass ein grosser Anteil der häufig benutzten Tags in georeferenzierten Flickr-Daten Ortsnamen entspricht. Die Stadt-Ebene ist dabei die Granularität, an die Menschen bei der Zuweisung von Ortsbeschreibung intuitiv am ehesten denken. Der Anteil generischer Begriffe, die den Stadtkern bezeichnen, ist meist marginal und die Mehrheit solcher Tags werden in Anbetracht der grossen Datenmenge nicht von vielen verschiedenen Benutzern verwendet. Vor allem in grösseren Städten ist die Verwendung von spezifischen Quartier- und Regionsnamen jedoch häufig. Die Beschaffenheit der abgeleiteten Footprints ist ein Hinweis dafür, dass der durchschnittliche Benutzer eine treffende und deutliche Vorstellung von spezifischen Orten hat und dass die Art und Genauigkeit mit der in Flickr Metadaten erstellt werden den Anforderungen für praktische Anwendungen sogar auf der Quartierebene genügt. Trotz der Verfügbarkeit einer riesigen Menge empirischer Daten machen die Auswertungen eine beträchtliche Ver-

zerrung durch einzelne Benutzer als auch hinsichtlich der räumlichen Verteilung deutlich, ein Umstand, dem bei der Verwendung solcher Daten unbedingt Rechnung getragen werden muss. Ist eine 'kritische Masse' von Tags und teilnehmenden Benutzern jedoch gegeben, so zeigt die Häufigkeit und die Qualität formaler und sprachlicher Ortsreferenzen in den Daten, dass diese geographisch wertvolle Information enthalten und geeignet sind für die systematische Erfassung von lokalem Ortswissen.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The advent of GPS-enabled phones and cameras, location based services such as Google Maps[1] or Google Earth[2], local search, the Geography Markup Language (GML), and facilities to geotag web content has turned the Internet into a key source of geographically related information (Erle et al., 2005). The burst in the use of the Internet and of consumer electronics came along with the emergence of social media sites on the Web 2.0, including wikis, MySpace[3], YouTube[4], or the photo-sharing platform Flickr[5]. A wide range of users do not only seek information but are actively participating in the creation and distribution of information (Lerman and Jones, 2006). They have become the major creators of geographical information, featuring locative aspects beyond the digital map data traditionally used in GIS (Erle et al., 2005). It is an issue of intensive research how to extract, mine, visualise and exploit this geographically relevant information (Boll et al., 2008).

Geographic Information Systems (GIS) have advanced from a primarily academic discipline to a mainstream technology. The increasing importance of geographic information in everyday life has brought a shift towards more psychological and social aspects in the traditionally technical field and GIScience has gained interest in systems that are able to handle common-sense geographic knowledge and human conceptualisations of locality (Egenhofer and Golledge, 1998). Humans acquire extensive spatial knowledge in the course of life, but there is a lot of vagueness inherent in the way we conceive and refer to geographic location. Instead of exact distances and coordinates, we employ vague spatial relations such as 'near the station' and talk about places such as London without being concerned about the precise nature of their boundaries (Montello et al., 2003). Other place names correspond to cognitive regions with lacking or inconsistent definitions; they are not legally

[1] http://maps.google.com/
[2] http://earth.google.com/
[3] http://www.myspace.com/
[4] http://www.youtube.com/
[5] http://www.flickr.com/

defined such as the 'Alps' or the 'West End' (Jones et al., 2008). Such common place names are the issue of vernacular geography, the geography of everyday space and language. It comprises a complex set of places at various scales, their often vague extents, and their multiple names with different meaning to different groups of people (Davies et al., 2008).

Despite the ubiquity of place names in every-day discourse, spatial search engines still cope poorly with vernacular place indications (Hill, 2006). Attention was therefore drawn to the automated interpretation of arbitrarily employed place names, which is also important to location-based services, map data providers, travel and emergency services, transport, and navigation (Davies et al., 2008). The issue involves geography and GIScience, as well as spatial cognition, linguistics and ontology, urban design and modelling, computer science and artificial intelligence.

1.2 Aims and objectives

The importance of understanding vagueness is widely acknowledged in GIScience today. There is a growing body of work focussed on the formal representation and computational implementation of objects with vague boundaries into GIS (Burrough and Frank, 1996). Other work suggested approaches to capture information on vaguely defined regions and presented techniques to approximate the regions' extents (Montello et al., 2003; Purves et al., 2005; Arampatzis et al., 2006; Grothe and Schaab, 2008).

This Master's project is not focussed on the problem of mathematical and computational formalisation of vague objects but on the question of how people think and communicate about geographic space and on the process of getting knowledge on people's collective understandings of places. The approach adopted takes advantage of the characteristics of an immense set of empirical data; the user-generated content on the photo-sharing platform Flickr. Flickr was chosen over other applications as it provides a database of global scope, which is easily accessible by an Application Programming Interface (API). Furthermore, at the time of writing, there were about ninety million georeferenced photos available on Flickr[6].

Despite the lack of pre-defined categories and restrictions in tagging systems, experiments have shown that user-employed keywords (tags) in Flickr exhibits patterns allowing for the extraction of valuable information (Rattenbury et al., 2007; Grothe and Schaab, 2008). Furthermore, the user-generated metadata is said to reflect people's perceptions, providing a novel, challenging opportunity to investigate distributed cognition and knowledge (Steels, 2006). Within the framework of this project, the aim is to take advantage of the abundance of absolute references between locality and user-assigned descriptions of place in natural language. In view of the availability of a global and multilingual database, the investigation of geographic terminology used for tagging covers German as well as English language areas. Georeferenced Flickr data is collected within the extent of different cities and evaluated in order to answer the following research questions.

[6] http://code.flickr.com/blog/2008/10/30/the-shape-of-alpha/, accessed 14th November 2008

- How do people describe urban places in terms of tags depending on language and culture region?

- Is user-generated metadata in online photo-collections suitable to capture vernacular geography?

- How can digital footprints of vernacular regions be modelled from georeferenced tags?

The primary objective is to explore how people commonly refer to geographic space, by analysing the kind of expressions that are used in terms of tags within different cities. The large-scale analysis will not include vague spatial relations such as 'near', but is restricted on vernacular places within the city environment. Cities are the main hub of social as well as commercial activity of the modern information society and a focus of geographic research as a whole (Knox and Pinch, 2000). In order to allow for an intercultural comparison, the investigation is focussed on vague concepts, respectively generic terms, used to refer to the urban core. From a practical point of view, the quantity of photos posted within central city neighbourhoods was assumed to be high enough to capture a distributed understanding of places.

Due to the chaotic nature of tagging, the benefits of user-created metadata are controversial. The second research question is aimed at an in-depth evaluation of the capabilities of social tagging systems for capturing collective geographic knowledge. The evaluation considers Flickr-specific characteristics of user-generated metadata associated with online photos but is, to some extent, intended as a case study of tagging systems in general. Thirdly, an approach to visualise vernacular places with knowledge gathered from georeferenced Flickr tags is developed. The problem comprises questions such as: How well does a method based on place semantics from Flickr work to derive digital footprints? What is the nature of the regions that result?

The publicly available API of Flickr enables an automated collection of an amount of data exceeding the capacities of manual browsing and allowing for a large-scale evaluation in order to answer the above research questions. Due to the lack of standard techniques in this context, the project was completed by adopting an explorative approach. The definition of further problems and the way to approach them was based on previous results. The analysis of tags corresponding to place names which occurred in the data samples mined from Flickr was completed by data-driven as well as theory-guided techniques. In order to visualise results, novel approaches for visualising tags as well as standard GIS techniques were adopted and revised. The results obtained from different methods, which were employed to analyse the geographic origin and the distribution of generic place tags, were cross-validated and the method developed to map digital footprints was evaluated by means of spatially well-defined public parks. Footprints derived for vernacular places were assessed by visual inspection.

1.3 Thesis outline

This thesis is organised as follows. Chapter 2 describes the theoretical background providing the motivations and assumptions underlying this thesis. First, the reasons for vagueness in the geographic domain and the implications of vague geographic objects on the formalisation process within GIS are considered. In particular, previous attempts to explore and model vague places are reviewed and the characteristics of common-sense representations of geographic space are disclosed in Section 2.1. The theoretical basis for the evaluation of vernacular terminology for the ill-defined city core is outlined in Section 2.2 and an overview of previous research on web-based tagging systems with a particular focus on Flickr is given in Section 2.3. In Section 2.4, the assumptions and research questions are further motivated.

The empirical basis of the study is described and the quality of the data collected from Flickr is examined in Chapter 3. An overview of the methods employed to explore georeferenced place tags is given in Chapter 4. A technique for the objective generation of vague footprints is evaluated in Section 4.2.2.

In Chapter 5, the results obtained from quantitative tag evaluation are presented with respect to both the characteristics of employed place indications and the culture-specific terminology for the ill-defined city core. In Section 5.3, vague footprints are derived and evaluated by visual inspection and qualitative comparison.

Chapter 6 provides a discussion considering the use of vernacular place names in the urban environment, the potential of Flickr data to explore collective understandings of geographic space, and an assessment of the suggested technique for footprint approximation. In Chapter 7, the most important accomplishments and results of the study are highlighted and consequent research problems are formulated.

Chapter 2

Background

The background of this thesis is based on the theoretical framework of a variety of academic fields. Not only GIScience and GIR, but also cognitive science and linguistics as well as urban geography and research in the field of tagging systems influence the fundamentals behind the assumptions and methods employed in the project. Related work is reviewed on the basis of a shared concern, expressed by the urban geographer Murphy (1972: 2) with respect to the central business district (CBD) as follows.

> "The CBD has no fence around it, no wall as there was around the city in Europe in the Middle Ages. You will never see a sign, "You are Entering the CBD", although there may be signs directing you to the city's downtown area. However, the district can be conceptualized and its position outlined on a map on the basis of this mental construct. How can this best be done?"

Based on the issue of vague geographic objects, the chapter as a whole provides an insight into the fundamental discussion about the configuration of space in the geographic discipline, ranging from the objective paradigm, aimed at full quantitative modelling, to the subjective framework, taking into account human views and actions.

2.1 GIS and vague geographies

2.1.1 Uncertainty in spatial information

Geographic Information Systems are tools aiding the collection, storage, manipulation, and display of spatial data. GIS provide a means for spatial analysis at geographical scales by representing reality in digital models. When using GIS for spatial analysis, we have to be aware of the fact that the object or field based models may involve a considerable amount of imperfection. GIScience has long focussed on the imperfection caused by poor data quality and error propagation in data processing. This uncertainty induced by observation bias,

measurement error, interpolation, generalisation, or classification is referred to as *inaccuracy*. *Imprecision* on the other hand, is a reference to the inexactness of the representation in terms of recorded decimals. Inaccuracy and imprecision relate to the imperfect state of our knowledge about an object's properties due to empirical shortcomings. They are typically treated by classical probability of conventional statistical analysis, assuming that there is some objective reality of boolean entities to which the represented objects can be compared (Burrough and McDonnell, 1998).

Uncertainty lying in the indeterminate nature of the mapped objects themselves is referred to as *fuzziness* or *vagueness* (Fisher, 1999). The 'sorites paradox' has frequently been applied in the context of GIScience to explain the notion of vagueness. It can be illustrated by the example of a skyscraper. If we look at a two-storey house, everyone will agree that this is not a skyscraper. By building another storey on top, we still do not obtain something we would refer to as a skyscraper. If we keep adding floors and continue our logical argument, the addition of a single storey would never turn the building into a skyscraper. By substituting the storeys and the skyscraper by rocks and a mountain, or buildings and a neighbourhood, we realise the implication of the logical paradox for the geographical domain (Couclelis, 2003).

In conclusion, one or a combination of the following reasons (adapted after Evans (2004) and Montello (2003)) account for different manifestations of uncertainty and vagueness within GIScience:

- Inaccuracy and imprecision (*epistemological vagueness*): Where the knowledge of a boundary is imprecise or inaccurate as it cannot be defined precisely due to empirical shortcomings, for instance, when relying on a satellite image with limited resolution.
- *Averaging vagueness*: Where boundaries vary with time or scale, for instance the transitional boundaries of the North Sea due to high and low tide and the respective representation on maps of different scales.
- *Multivariate vagueness*: Where alternative combinations of variables are possible for categorisation. An example is the discrete categorisation of soil types into prototypes of a taxonomy.
- *Contested vagueness*: Where there is disagreement about the course of a boundary. An example is the different conception of the boundaries of Europe, whether based on natural or political factors, or the delineation of the same place by different individuals.
- *Conceptual vagueness*: Where the underlying concept is per se ill-defined. An example of the conceptual vagueness problem is defining what a mountain is or how far a neighbourhood extends.

Within the scope of this thesis we are mainly interested in the conceptual and contested terms of vagueness. The subsequent sections provide an in-depth review on the nature of such vague places, their boundaries, and their formalisation within GIScience.

2.1.2 Vague regions and boundaries

Unlike manipulable table-top objects such as books or computers, the category and extension of the majority of large-scale geographic entities is hard to define; they are said to be vague (Frank, 1996). Or, as expressed by Couclelis (1996: 48): "Often the transition from one geographic entity to another is smooth and continuous, so that any boundary between them is conventional rather than empirically real." Vagueness in the geographical world comprises considerations about what a mountain is and where it begins, how far the Australian outback extends, or where to mark the borders of Central London. The majority of natural geographic phenomena like vegetation zones or soil types are spatially ill-defined. A review of their treatment in GIS models is given by Burrough (1996). We will here concentrate on the second type of vague geographic entities: human conceptions such as aboriginal territories or urban neighbourhoods.

There is a strong interrelation between regions and their boundaries. Vague in a spatial sense means, that the boundary around an entity is not basically one-dimensional but a two-dimensional zone of gradual transition (Montello, 2003). A twofold typology for the nature of boundaries with consequences for the conceptual assumptions in GIScience was proposed by Smith (1995): *Fiat* boundaries are social creations resulting from human conventions and conceptions, while *bona fide* boundaries exist independently of human conventions and cognition. The first include both the well-defined demarcations of legal borders and more vaguely drawn regions such as 'Middle Europe'. As individual and social fiats do often not coincide with spatial discontinuities, they are typically subject to vagueness in people's minds. For a cross-country hiker in the Alps, for instance, it might be hard to tell on which side of the Swiss-Italian border he is actually located. Boundaries manifested in physical discontinuities of the underlying reality are referred to as *bona fide* boundaries. The Rhine delimiting the two landmasses of Switzerland and Germany is an example of a fiat border type marked by a perceivable bona fide boundary.

Montello (2003) proposed a taxonomy for anthropo-geographic regions based on process and content, consisting of four types with varying degree of boundary vagueness: administrative, thematic, functional, and cognitive regions. Both thematic regions, defined by the occurrence of shared characteristics, as well as functional regions, formed by patterns of interaction and flows between places, tend to be vague. Administrative regions are established by legal and political action. They are defined by precise coordinates and well-defined boundaries but are not necessarily manifested in physical space. Cognitive regions are formed by people's casual conceptualisations and are "typically fundamentally vague, with every crisp representation a fiction to some extent." (Montello, 2003: 180). As Montello (2003) acknowledges, the distinction between the four types is in practice somewhat blurred, as the production of regions is context-dependent. While 'Sheffield', for instance, is a well established administrative region in terms of a borough, it is also a cognitive place in humans' minds. People talking about Sheffield might rather refer to Sheffield city or the city centre than Sheffield borough as a political unit. Even though the residents of a city have a general impression of its legal boundary, most of them will not be able to precisely trace it on a map. It is turned into a vague region due to the limitations of human observation (Fisher, 1996). Hence, regardless of the ontological and empirical

characteristics, whether a place appears crisp or well-defined is determined by political, physical, social, and cognitive processes on the one hand and the mode of observation and representation on the other hand (Couclelis, 1996).

Cognitive conceptions about regions shared among many members of a culture are referred to as *vernacular regions* by geographers. They are a subtype of cognitive regions, corresponding to a collective conception about a place, an associated name and an approximate extent (Montello, 2003). Vernacular places are a very persistent, but not a static component of a culture and society (Hastings, 2008). Vernacular geography or 'how we speak about places' is strongly connected to people's beliefs about a particular place. This has been referred to as 'sense of a place' in human geography, meaning that the extent of such a place is defined by the characteristics and the experiences made possible by it (Tanasescu and Domigue, 2008). That such places typically have vague boundaries in people's minds will further be established in the next section.

2.1.3 Common-sense knowledge of geographic space

Humans acquire extensive knowledge about geographic space in the course of life. The belief that this knowledge, and the natural language used to express it, need to be accounted for to make geographic information technology more efficient explains the growth of interest of GIS in the problems and findings of cognitive science as well as linguistics (Egenhofer and Golledge, 1998). Geography is interested in the process of acquiring knowledge of large-scale geographic space, the spatial configurations that are beyond our immediate sensory experience. Decades of research in various disciplines such as environmental and cognitive psychology, artificial intelligence, and ontology have contributed to the understanding of how we perceive, categorise, and apply geographic information and how we communicate geographic knowledge (Montello and Freundschuh, 2005).

Influential for the understanding of how people think about geographic space is literature concerning cognitive and linguistic category theory. Related studies (Rosch, 1978; Lakoff, 1987) reveal that humans make sense of their experiences through the cognitive process of categorising of 'what is out there' by making use of idealised cognitive models (ICM). These allow for fuzzy and overlapping categories. The problem of gradual transition between categories is inherent to the process of categorising reality into taxonomies and is reflected in the graduality of natural language (Fisher, 1999). Vertically, the categorisation is hierarchical and supports the notion of basic-level categories. The basic-level is at the middle of the taxonomic hierarchy and characterised by the fact that it requires the least cognitive effort and that no increase in knowledge is achieved by further specialisation. Basic-level categories represent the default level of abstraction for reasoning and discourse (Lakoff, 1987). The levels applicable to 'Lassie' from the well-known television series, for instance, range from 'animal' to 'rough collie' but most people would consider most useful to simply describe her as a 'dog'.

Geographical entities are no exception in terms of how meaning is associated to them. As remarked in Section 2.1.2, the occurrence of graded membership and unclear categories is

particularly true for geographic objects, in terms of category as well as in terms of spatial extension (Frank, 1996). While it can be understood what the concept of 'town' or 'city' comprises, it is difficult to establish the differentiation between the two. Also, the categorisation enabled by these concepts does not mean that they allow for the identification of the boundaries of specific cities or towns (Ferrari, 1996). The fact that human categorisation of reality is reflected in linguistic concepts was frequently used to study our understanding of geographic space (Ferrari, 1996). McGranaghan (1990), for instance, explored how humans conceptualise geographic entities of the physical environment by studying textual place descriptions in large herbarium records. Smith and Mark (2001) investigated what people consider to be the most typical examples of geographic entities in order to establish if there is a basic level of categories in the geographic domain. Comparative studies (Mark and Turk, 2003; Ferrari, 1996), also based on entities of the natural geographic environment, have shown that the boundaries of linguistic categories are rarely congruent between different languages representing different culture regions.

The framework of mental categorisation has been applied to the experience of geographic space amongst others by Gale and Golledge (1982), McNamara (1986) and Mark et al. (1999). Decades of research have revealed that the hierarchical structuring of space is essential to the development of spatial knowledge. The acquisition of spatial knowledge is basically constituted by the process of low-level sensational perception, followed by high-level cognitive categorisation. While the mental structuring of space is "largely culturally universal" (Montello, 1995), it is guided by culturally and socially determined categories and concepts. Knowledge about the large-scale geographic environment may be acquired from direct experience and observation, referred to as primary learning. Secondary learning includes learning from graphical representations, most importantly maps, as well as spoken and written language. To get an idea of a coherent whole, we integrate knowledge about fragments of geographic space as moving about, or by completing our knowledge through plans and maps (Kitchen and Blades, 2002). The requirement of piecing various parts together to gain a complete integration of the environment is commonly seen as a key factor in the process (Montello, 1998). Insight into how people structure and organise their understanding of geographic space has been derived from the investigation of people's 'cognitive maps'. The term was first used in a paper by Tolman in 1948 to describe the representation of spatial information in human memory (Gale and Golledge, 1982). From the study of cognitive maps, there is substantial evidence that humans think about geographic space in terms of entities and the relations between those rather than in terms of coordinates and exact distances. Mental maps are primarily based on qualitative and topological relations rather than quantitative measures, while the length of residency, urban experience, and navigational abilities considerably influence the cognitive structure (Couclelis, 1996). Due to the topological and hierarchical nature of our geographic knowledge, we are able to retain spatial relationships and the hierarchy of places (Hill, 2006). Hence, we are more likely to reproduce that Soho is a part of the City of Westminster and that Chinatown lies within Soho, than to tell the exact distance from Trafalgar Square to the Piccadilly Circus.

The main properties of cognitively derived regions is their hierarchical structure and their fuzziness in terms of boundaries (Gale and Golledge, 1982; Montello, 2001; Hirtle, 2003).

Despite the vague nature of human representations of the environment, investigations of cognitive maps have revealed that there is a considerable amount of consensus between the structure of individual cognitions. Geographic entities of urban space, for instance, tend to have collective definitions, which are based on cultural, social, and historical conventions and which are marked in the minds of the residents (Ferrari, 1996). Various aspects of culture, most importantly language, provide the background ensuring the significant agreement on the concepts applied by different individuals (Gale and Golledge, 1982). "The body of knowledge that people have about the surrounding geographic world" has been termed common-sense or 'naïve' geography by Egenhofer and Mark (1995: 4). In view of the growing community of (non-expert) GIS users, it has been accentuated that geographic information technologies need to account for such vague and multiple conceptualisation of geographic space (Egenhofer and Golledge, 1998; Montello and Freundschuh, 1995; Smith and Mark, 2001; Hirtle, 2003). We typically refer to location in terms of hierarchically structured and often ill-defined places. Place names are a crucial concept in communicating geographically relevant information on day-to-day basis, while location indication in traditional GIS heavily relies on the specification of geographic coordinates (Hill, 2006). The consideration of the cognitive and linguistic dimension should not only include the interfaces, applications, and tools, but also the data, as well as the representations. Montello (2001) and Hirtle (2003), for instance, argued that environmental knowledge is not well represented in metric geometry but rather as a set of fuzzy categories.

2.1.4 Vernacular geography and GIR

Information Retrieval (IR) is a well-established discipline dealing with the extraction of relevant information from unstructured collections on the basis of a query. IR, as we are all familiar with from web search engines like Google, is primarily based on text-dependent methods in combination with some type of ranking, yielding a result list of information objects in decreasing order of relevance (Purves and Jones, 2006). As the majority of information is implicitly or explicitly related to some location on Earth and, as established above, people think about geographic location in terms of places rather than exact coordinates and extents, queries submitted to information systems are likely to contain some notion of a place name. While even 70% of text documents contain references to named places (Hill, 2006), a classification by Sanderson and Kohler (2004) revealed that almost one fifth of analysed web queries were geographically related. Nearly 80% thereof where specified by the use of a place name. A more recent analysis of about 36 million queries of the AOL query trace found that about 13% of the queries contained some kind of a place indication. The authors also showed that queries at different levels of granularity cover different information needs (Gan et al., 2008).

Studies have revealed that text-based search does not cope satisfactorily with place relationships. It is sensitive to spelling, language, and ambiguity instead of accounting for semantic information (Hill, 2006). Motivated by the automatic georeferencing of text documents and spatial browsing in digital libraries, information retrieval has been extended to Geographic Information Retrieval (GIR) by Larson (1995). It has later been defined

by Purves and Jones (2006: 375) as "the provision of facilities to retrieve and relevance rank documents or other resources from an unstructured or partially structured collection on the basis of queries specifying both theme and geographic scope." GIR has received growing scientific as well as commercial interest in recent years. Today, research in the field includes the establishment of geographical ontologies, geoparsing and spatial indexing of documents, disambiguation of place indications, and the development of relevance ranking algorithms (Purves and Jones, 2007). In the following, we will focus on the aspects of GIR related to vernacular place names.

The identification of place indications in text or on webpages, referred to as *geoparsing*, and their resolution into coordinates, known as *geocoding*, is usually performed by gazetteer lookup, a key component of geographic information services (Larson, 1995). Digital gazetteers are hierarchically structured lists of named places, relating between a location represented by a textual label and its formal geospatial location in terms of coordinates (Hill, 2006). Examples of major online gazetteers are the Getty Thesaurus of Geographic Names[1] (GTN) and the gazetteer service of the Alexandria Digital Library[2] (ADL). According to a standard developed by the ADL project, a gazetteer entry should at least include the core elements of (1) a place name, (2) its footprint represented by coordinates, and (3) a place type designation, for instance assigned according to the Feature Type Thesaurus associated with the ADL. Gazetteers may include all kind of additional information such as descriptions, the temporal dimension, alternative spellings, and former and colloquial names (Hill et al., 1999).

Footprints associate named places to their geographic location on Earth. In standard gazetteers they are usually stored as single points, typically at the centroid of the location. In more sophisticated systems, they may also be represented in terms of linear features, bounding boxes, or detailed polygons. GIR is based on the comparison between geographic queries and the footprints in a gazetteer, yielding a degree of relevance of potential information objects. Relevance in the domain of IR and GIR is defined in terms of the usefulness of a response in relation to the user's information needs. The performance is a trade-off between recall (portion of relevant objects that are retrieved from the total number of relevant objects in the collection) and precision (portion of relevant objects among the retrieved objects) (Hill, 2006).

As shown in Sections 2.1.2 and 2.1.3, places are a crucial aspects of a particular culture and language, reflecting both official regulations and societies' local identities. Place name information for gazetteers is typically local cultural knowledge and authorities in charge of the definition of officially approved toponyms provide the main source for digital gazetteers. Other examples of the countless contributors to gazetteer data include governmental and private publishers of maps, planning agencies, property ownership and constructing registers, and library collections (Hill, 2006). Currently, most GIR systems only work satisfactorily with respect to place names belonging to the category of well-defined, administrative units or physical and cultural features being used in official topographic maps. Gazetteers generally do not store entries for vernacular places which are

[1] http://www.getty.edu/research/conducting_research/vocabularies/tgn/, accessed 15th Sep 2008
[2] http://www.alexandria.ucsb.edu, accessed 15th Sep 2008

often used in everyday discourse and are typically subject to vagueness. Hence, attention was recently drawn to the use and automated interpretation of arbitrarily employed place names, which are challenging and difficult to process by computer systems (Hill, 2006), as well as to the development of standards towards the incorporation of vague places at different levels of regional hierarchy to populate gazetteers (Goodchild et al., 1998). The importance of the implementation of fuzzy footprints into geographic search systems was also mentioned by Purves and Jones (2007), who identify the effective handling of common sense spatial knowledge and uncertain spatial relations like 'near' and 'outside' as one of the key requirements for GIR.

2.1.5 Sources of knowledge on vague places

As many vernacular place names lack formal definition, or the general perception of a place might differ from the official definition, the question arises how to acquire knowledge about the nature of such regions (Twaroch et al., 2008). The literature review on the subject revealed that knowledge about the location and extent of vague regions and vernacular place names have been acquired from a wide variety of sources. Early attempts made use of descriptions in written language to acquire factual information of people's impressions about a place at a particular epoch. Byrkit (1992) has interpreted textual publications mentioning certain places to be internal or external with respect to the ill-defined American Southwest. The author found that the Southwest is defined in many terms, for instance at the political, historical, cultural, mythic, physiographic, and bureaucratic level, before delimiting the region at specific meridians and circles of latitude by classical geographic regionalisation. Llyod (1976) has studied individual geographic awareness with respect to small-scale regions and was able to identify and outline different places in the city of Boston from novels.

The first attempt considering people's cognition of ill-defined places with a background in GIScience was carried out by Montello et al. (2003). The authors investigated pedestrian's perception of downtown Santa Barbara by having them outline their conception of the region on a base map. A probabilistic model of the downtown was derived by adding up the binary maps from the participants. As acknowledged by the authors, a drawback of the method is the bias induced by the size of the base map. The technique has therefore been adapted to asking pedestrians if they placed a landmark inside, outside, or on the border of the city centre of Sheffield. Mansbridge (2005) found that the average human cognition of the city centre comprised a smaller area than the various inconsistent official definitions. Human subject tests and interviews are a very powerful means to investigate people's conceptualisation of single regions, but suffer from a scalability problem for the purpose of populating gazetteers (Twaroch et al., 2008).

A web-based GIS technique to capture people's ideas of fuzzy places was implemented by Waters and Evans (2003). Users were asked to identify high crime areas in the city of Leeds by means of a spray can tool, allowing them to draw areas of varying density and fuzziness. The recorded drawings were converted into an aggregated map of combined density surfaces. Lam et al. (2002) describe a method depending on the text on maps

to derive footprints for neighbourhoods in the city of Los Angeles. The textual labels were treated as the centroids and the mean distances between neighbouring centroids were taken as a reference for the radii of circular footprints representing the neighbourhoods. The resulting pattern of variably sized circles accounts for the fact that neighbourhoods are both not space-filling and sometimes overlapping.

Evidently, the steadily growing web is a major source of knowledge about vernacular places that has frequently been taken advantage of. Purves et al. (2005) first successfully modelled vague regions by using an approach which is based on the assumption that vague place names and well-defined toponyms will co-occur on websites. Simple or trigger-phrase queries containing the vague target region are submitted to search engines like Google and the highest ranked results are searched for toponyms by automated text mining. The geoparsing process resolves place names overlapping with names of non-geographic objects and proper nouns (referent class ambiguity). Identified toponyms are geocoded with coordinates through gazetteer lookup. If more than one possible match is found in the gazetteer (referent ambiguity) a disambiguation technique is applied in order to get a set of points likely to be located within the vague region. The resolution of the possible two-way ambiguity of toponyms is one of the main challenges of the approach. Also, web entries are biased towards places with higher population density or popularity (Jones et al., 2008).

Twaroch et al. (2008) collected knowledge about vernacular places at the neighbourhood level from absolute references between place names and location in users' posts on a social trading website[3], and in Google community maps. The mapping of the derived regions revealed that some of the footprints were spatially congruent with their official counterparts, while others were not. The project most closely related to the approach pursued in this thesis was presented by Grothe and Schaab (2008), who successfully modelled colloquial conceptualisations of the Alps, the Black Forest, and the Rocky Mountains from georeferenced tags in Flickr.

2.1.6 Models, formalisation, and the delineation of boundaries

Traditional GIS represent geographic objects either by sharply delineated, homogeneous objects or continuous fields. The nature of vague geographic objects is not suitably represented in the conventional object and field dichotomy (Burrough and McDonnell, 1998). Both the vector and the raster based model have been adapted for the presentation of vagueness. Other approaches benefit from the fact that qualitative topological relations are not affected by the fuzziness of boundaries and use topological reasoning to resolve the issue (Burrough and Frank, 1996). Even though vague regions do by definition not have a single, precise boundary, it depends on the application if the approximation of crisp boundaries is appropriate. It is indisputable that for certain purposes, the delineation of a single, well-defined boundary is useful or necessary (Jones et al., 2008; Davies et al., 2008). In this section, several approaches to formalise and model vague places as well as methods to generate sharp boundaries for vague regions are reviewed.

[3] www.gumtree.com

Zadeh (1965) has first introduced the idea of 'fuzzy sets' to deal with vague concepts. Fuzzy sets, which allow for overlapping and indefinite memberships in the classification process, are not a probabilistic but possibilistic approach (Fisher, 1996). In a spatial context, the fuzzy membership function expresses partial affiliation by a value gradually fading from 1 at the centre of the element towards 0 at locations outside the set. Fuzzy sets have traditionally been used in geography to analyse physical phenomena such as land use, soil classification, and pollution mapping (Burrough and McDonnell, 1998). Recently, Schockaert and Cock (2007) have taken advantage of fuzzy set theory to derive fuzzy membership footprints of vernacular places from Yahoo! local[4] data. The approach was tested for neighbourhoods in the city of Seattle.

Cohn and Gotts (1996) suggested an often-cited system for representing vague regions based on rough sets. In this system, referred to as the 'egg-yolk' model, the yellow of an egg corresponds to the assured core of a region and the egg white to the uncertain zone of transition. Vague places are approximated by two (or more) concentric subregions, each indicating the assumed degree of membership. From this model, topological reasoning of places with indeterminate boundaries is derived by using a framework for crisp objects. Another topological approach was presented by (Vogele et al., 2003), who approximated vague places by their qualitative relation to officially defined regions. The lower approximation is given by the official regions which are definitely contained by the vague place, while the upper approximation additionally comprises the administrative regions overlapping with the place under consideration.

Different approaches have been presented to model the extent of ill-defined regions based on georeferenced candidate points. Footprints of vague regions intended for the use with gazetteers are delineated by Voronoi diagrams of administrative point locations known to be located inside or outside the target region (Alani et al., 2001). Arampatzis et al. (2006) use Delauny triangulation on inside and outside-classified points from web mining and employ *α-shape* or *recolouring methods* for a refinement of the boundaries. The method was tested and evaluated by approximating the boundaries of the four regions of Wales, the Midlands, the South East, and East Anglia in the UK.

Several authors modelled the confidence of a point location being inside the vague region by using spatial density estimations. If crisp boundaries are required, a point density might be selected to threshold the density surface. Purves et al. (2005) and Jones et al. (2008) compute kernel density estimation (KDE) by weighting the candidate points from prior web search by term frequency (the total number of occurrence of a place name in the retrieved documents) or document frequency (the number of documents a place name is found). In this way they successfully generated models for large-scale regions such as Wales, the Swiss 'Mittelland', and the Scottish Highland. Twaroch et al. (2008) estimated kernel density surfaces from georeferenced points in community websites to model vernacular places in the city environment. The technique was enhanced and benchmarked by Henrich and Lüdecke (2008) to resolve footprints for web queries with unknown locators in real time. Also Grothe and Schaab (2008) used KDE to estimate footprints of large-scale geographic regions from georeferenced Flickr tags.

[4]http://local.yahoo.com/

2.2 The city core

According to the main objective of regionalisation within geography as a whole, urban geography has for long been interested in the internal structure and differentiation of urban space, which have been analysed from a variety of perspectives (Knox and Pinch, 2000). Below, some of the major approaches of urban geography are briefly reviewed with regard to their application to the city centre, the vague region in the focus of this thesis. The second subsection describes the origin of geographic terminology to describe the ill-defined city core as well as actual applications of such generic terms.

2.2.1 Urban geography and the city core

The ecological approach

By introducing the 'Concentric Zone Model', the sociologist Ernest Burgess provided the theoretical foundation of the ecological approach. The qualitative model of ecological change was devised with Chicago of the 1920s in mind, and structured the city into a set of five concentric zones of decreasing intensity of land use. These spatially fixed areas are defined by population and functional shifts induced by land market competition and conflicts between different social groups. In Burgess' model, the zones expand outwards from a core termed central business district (CBD), which was identified as the main hub of commercial activity. It spreads into the 'zone of transition', a mixed-use area occupied by wholesaling, light industry, and dense housing, accommodating the urban underclass (King and Golledge, 1978). The concentric zone hypothesis was later amended by Hoyt's 'Sector Model' and the 'Multiple Nuclei Model' of Harris and Ullman, which accounts for primary centres (e.g. London) or secondary centres (e.g. Chicago) of the multi-nucleated metropolis. Even though Burgess' model has soon been criticised for its biotic analogy, traditional social ecology was carried forward by *factorial* and *social area analysis*, two approaches aiming at the quantitative determination and classification of socio-economically homogeneous neighbourhoods.

The functional approach

The functional approach is primarily concerned with the analysis of functional entities of urban space as well as with the historic dimension of the CBD formation. The functional change of the city core since the 1850s was marked by the substantial loss of residential population, the augmentation of commerce and employment, as well as the proliferation of private traffic. The concentration of major business activity at the geographic core reflects the advantage of proximity that is highly characteristic for the financial sector. Apart from bank headquarters and stock exchange, the main functions nowadays hosted by the city centre are insurances, department stores, business hotels, media and newspaper companies as well as theatres, entertainment centres, and restaurants (Gaebe, 2004).

The functional approach adopted quantitative techniques embedded in the neo-classical paradigms of the 'quantitative revolution' of the 1960s, when geography as a discipline was striving for more scientific systematisation and respectability (Heineberg, 2000). A prominent example of the analysis of geographical structures by statistical measurements was the concern for an exact and generally applicable delimitation method for the central business area, as opposed to locally understood boundaries. The variables measured in order to delineate the CBD included the elaborate classification and mapping of building occupancy, building type and height ('skyscraper index') as well as the measurement of traffic flow, volume of trade, and population density (Murphy, 1972). Based on the study of nine moderate-sized American cities, Murphy (1972) suggested the central business index method of delimitation (CBI). The method is based on a (arbitrary) distinction between non-central and central business activity and the calculation of two critical ratios by block units. As Heineberg (2000) states, the delimitation methods relying on quantitative measurements are to some extent problematic due to the gradual change of land use in urban space. Also Murphy (1972) acknowledges, that the calculated delimitations are believed to be fair approximations as the CBD boundary is a convention rather than a reality and a zone rather than a line on the map. He discusses a possible extension of the method to identify a core and a frame of the CBD by the further distinction of business activities. Even though the method was presented long before the age of GIS and the discussion about the computational formalisation of vague regions, the suggestion roughly corresponds to the 'egg-yolk' model later proposed by Cohn and Gotts (1996). Recently, Thurstain-Godwin and Unwin (2000) have presented a statistical technique for the robust and universal measurement of town centredness in the United Kingdom (UK). Composite KDE is derived by map overlay of data per post code unit, representing the four key factors of centrality, namely economy, constructional density, diversity of use, and visitor attractions. The GIS-based approach yields continuous density surfaces with the option of peak thresholding to derive crisp regions representing the town centres.

The behavioural approach

Taking into account people's cognition and evaluation of urban space, the behavioural approach was a direct reaction to the normative assumptions behind quantitative geography (Knox and Pinch, 2000). It is based on the believe that "in a sense the city is what people think it is" (King and Golledge, 1978: 4). The basics of the cognitive approach are discussed in Section 2.1.3. In his seminal studies Lynch (1960), for instance, considered the function of a couple of elements appearing in residents' cognitive maps with respect to how people divide a city into different districts. Lynch identified districts as two-dimensional subsets of urban space, which, depending on the external appearance of the city, are more or less essential in the image of the residents. People establish districts by a characteristic combination of components in the urban environment, such as form, structure, street patterns, and land use as well as type and condition of buildings. Districts are bordered by imaginary or real edges. While some edges, such as rivers, motorways, and train tracks, have a strong effect, as they are hard to cross or well distinguishable from a distance, such as skyscrapers against a park, other 'edges' correspond to smooth

transitions. Whether a boundary in urban space appears well-defined and doubtless, vague and extendable or is missing at all, is related to the degree of its physical manifestation in the environment (Lynch, 1960). This distinction can be compared to the cognitively motivated, i.e. fiat boundaries by Smith (1995), which might or might not be manifested by bona fide boundaries in physical space.

Lynch (1960) developed his framework of elements by comparing the external appearance of the cities of Boston, New Jersey, and Los Angeles to the images held by their inhabitants. With respect to the city centres, he found that the centre of Boston was laid out clearly in reality and in mind, having the Charles River as a distinct boundary on three sides. New Jersey was, due to careless development, considered to have not only one but four or five city centres – or rather none. Also the image of the city centre of Los Angeles was found to be fuzzy. Generally, the Broadway was designated to be the centre but did not have the correspondent functional meaning to the inhabitants. Due to advanced decentralisation the central area was only called downtown out of habit (Lynch, 1960).

The cultural-genetic approach

The aim of the cultural-genetic approach is to describe cities from different culture regions and to develop regional models of city types. It is based on the assumption that due to shared historical, cultural, and political influence, the similarity between cities from the same culture region is more significant than the diverseness between them. (Heineberg, 2000; Gaebe, 2004). In view of a data source allowing for intercultural comparison in this project, some insights about regional city types and their cores are provided. They demonstrate how people actually perceive and identify the city centre in different culture regions and explain the emergence of diverse generic terms used to refer to the urban core, even within the Anglo-Saxon language area.

Having a long historical background, the internal differentiation of European cities is more diverse than the concentric structure of the typical American city (Hofmeister, 1996). The steady incorporation of neighbouring municipalities into the core city evoked complex patterns of hierarchical settlement centres (Gaebe, 2004). Although fortifications have usually long been abolished, the main feature conferring identity and attractiveness to the cities are the historic buildings, churches, and towers situated in the compactness of the old town. The extensive growth of larger cities in the course of the nineteenth century resulted in a peculiar development of the East- and the Westend and in the formation of a central business core. In Europe, the process of functional differentiation generally rooted in the old town but expanded onward into the adjacent upper class residential districts and the railway station area, with the ‘station road’ as the major axis of city enlargement (Hofmeister, 1996). Nonetheless, multi-storey buildings are rare landmarks in the silhouette of European cities and located towards the fringe, as in the Docklands of London. In the course of increasing suburbanisation, city centres across Europe have lost residents and retail business to the periphery, but the complete removal of retail outlets and higher income groups from the central city was prevented. Since the 1980s, the differentiation of novel lifestyles and the ongoing process of gentrification have added to the desirability of

inner-city neighbourhoods due to their adjacency to attractive employment and formation opportunities, cultural venues, and leisure time facilities (Gaebe, 2004). To summarise, cities in Europe are nowadays characterised by historic constructions in the old town, renewal and gentrification of inner-city neighbourhoods, and ongoing suburbanisation with simultaneously politically, economically and socially dominant urban cores.

A specific variation of the European city is the British settlement, as neo-liberal authorisation policies of the 1980s have lead to more extensively spread agglomerations (Gaebe, 2004) and people's attitude is somewhat more anti-urban than in Continental Europe (Hofmeister, 1996). Particularly the City of London has largely lost its residential and supply function in the course of time. While it had about 200'000 inhabitants in 1700, it has turned into a major business district, home to only 5'000 permanent residents today. Another characteristic of large British cities is the more pronounced segregation and the formation of ethnic districts (Gaebe, 2004).

Except for some cities featuring colonial buildings, the typical North American city lacks a historic town in the old world sense (Gaebe, 2004). As the business core was often constrained by the presence of dense industrial districts, the increasing demand for office space in the late nineteenth century was supplied by the displacement of residential estates and the construction of multi-storey office buildings. Skyscrapers spread from New York and Chicago to literally every larger North American city in the late 19th century and were soon considered as the main symbol of the prosperous economy and the American way of life. The negative ecological effects of the proliferation and dependence of the automobile was one of the main reasons for the multi-causal decline of the American downtown, which reached its trough by the end of the 1960s (Fogelson, 2001). The relative importance of the downtown decreased dramatically due to the substantial loss of retail shopping, well-funded residents, and office space to the wider metropolitan area. 'Edge cities', sometimes called 'suburban downtowns', along strategic traffic nodes in the spreading agglomeration began to cover all functions of emergent cities. Due to the removal of many middle class inhabitants, the zone of transition at the fringe of the downtown was left neglected and decayed, becoming a main hub for social problems. Waves of immigrants settling in the heavily segregated inner-city neighbourhoods accounted for the emergence of common neighbourhood names like 'Chinatown' or 'Little Italy' (Gaebe, 2004). To summarise, the contemporary North American city core is characterised by comparably little residential population and retail outlets, the extreme clustering of high-end business activity and facilities, and the striking prevalence of skyscrapers. Apart from skyscrapers, distinct landmarks of the city core are large sports and convention venues, luxury hotels and long, orthogonal streets but no churches or squares such as in Europe (Gaebe, 2004).

Except for the historic core of Sydney, the Australian urban settlement is younger than 200 years in age (Hofmeister, 1996). The city core in Australia is characterised by clustered skyscrapers at the commercial centre, orthogonal street patterns, and the presence of extensive public parks and gardens (Gaebe, 2004). Although the appearance of Australian cities is in many ways similar to the American urban landscape, the concentration of employment towards the CBD, associated with masses of commuters, motorised traffic, and the subsequent decay of inner-city neighbourhoods is much less pronounced than in

the United States (Hofmeister, 1996). Additionally, recent tourist demands and the gentrification of central neighbourhoods assist the prosperity of Australian city cores (Gaebe, 2004).

2.2.2 City core terminology in different language areas

This section provides background information for the evaluation of place tags and establishes a link between the above considerations about the diverse characteristics of cities and the development of specific terminologies. Obviously, there is a variation of vocabulary between German and English language usage, but as the terms employed to describe the environment are closely linked to the evaluation thereof (Rapoport, 1976), we can also expect a variation between English language areas, namely between the British, American, and Australian subregions.

In towns of the German speaking world, the 'Altstadt' (old town) denominates the dense, formerly fortified historic part of a city (Gaebe, 2004). A variety of terms exist for the wider city core, which are subject to semantic overlap and inconsistent usage. The 'Zentrum' or 'Stadtzentrum', corresponding to the English expression 'centre', is well established by the geographic concept of *centrality*. It stems from 'centrum' for 'focus' or 'midpoint' in Latin and is found in all idiomatic derivatives thereof. As a colloquial expression, it is mainly used metaphorically to refer to the geographic core of settlements (Juchelka, 2001). The central business core of major settlements is referred to as 'City' and the respective formation of a differentiated district as 'Citybildung'. This terminology is adopted from the City of London, where a condensed financial sector already formed in the middle of the 18th century. The concept of the City was primarily defined from a functional as well as physiognomic point of view in urban research, but is today widely established in colloquial German linguistic usage (Heineberg, 2000). In the German sense of the word, the City is further differentiated into the 'Stadtkern' (city core), including the historic and the economic centre, and the 'Citymantel', the transitional zone surrounding the dual core (Heineberg, 2000). 'Innenstadt' is used to describe the city core and the adjacent neighbourhoods characterised by dense housing as opposed to the more loosely populated outskirts of the city (Juchelka, 2001).

In Great Britain and some of the Commonwealth countries, people would commonly use 'city centre' (Heineberg, 2000) or 'central area' Murphy (1972) to refer to the city core. Contributors on wikidepia[5] state that in Australia, South Africa, Canada, New Zealand, the UK as well as the New York area, city centre is often shortened to 'city' or simply referred to as 'town', for instance in the phrase 'going into town'. In the special case of London, 'the City' usually means the financial district in the City of London rather than any other central part of the Greater London Area (Heineberg, 2000). While the city centre is still considered as a mostly desirable and attractive location, the 'city center' in the United States is rather negatively connoted. The 'central city' is an expression for the municipality in the densely populated centre of larger metropolitan areas in the United

[5]http://en.wikipedia.org/wiki/Central_business_district, accessed 24th October 2008

States, were it is officially defined as a functional unit (Law, 1988). It is closely related to the 'inner-city', which is less technically defined (Caves, 2005) and usually employed in an evocative way to refer to the areas of poverty surrounding the centre, implying a socio-cultural negative connotation. It may also be used for the prosperous part of the centre and the surrounding neighbourhoods as a whole (Law, 1988). In this sense, inner-city is the more common expression than central city outside of the United States (Caves, 2005).

The term 'downtown' originated in the city of New York of the nineteenth century, when residents began to distinguish downtown, uptown, and midtown sections of Manhattan, according to their cardinal location. Lower Manhattan or downtown turned into the major centre of financial, wholesale, and retail commerce and was clearly distinct from residential upper Manhattan in people's minds. Even though it originated as a place name in New York, downtown lost its original geographical meaning and evolved into a generic expression that was soon applied to the city centre of every larger city throughout Northern America. Due to its specific characteristics it was thought to be uniquely American by that time and was easy to locate for everyone. Nonetheless, the downtown is hard to define as a place, as it does not exist legally and is typically transversal to the boundaries of the governmental wards (Fogelson, 2001). The concept did not remain a big-city phenomenon but was later used for the main shopping mall and its adjacencies of literally every American town (Caves, 2005) or even synonymous to the central city of an entire metropolitan area (Murphy, 1972). For a long time, it was practically unknown in Great Britain and Continental Europe, but in the age of globalised culture, media, and language, downtown is said to have spread over the globe (Fogelson, 2001).

'Central business district' or CBD was established by Burgess who referred, about twenty years after the term had first emerged in an American newspaper, to the innermost region of the Concentric Zone Model accordingly. The term caught on in the 1920s and was widely used for the vague region of concentrated economic activity in the city about two decades later. As for downtown, the CBD was neither legally defined nor politically formalised and its boundaries subject to constant change (Fogelson, 2001). Nonetheless, residents were "likely to know in a general way the location of the district in their city and to have a rough idea of its extent" (Murphy, 1972: 1). The functional sense of the word had early replaced the locational meaning and the concept was soon established in urban geography (Fogelson, 2001). According to Murphy (1972) the expression had not been used colloquially a few decades earlier, while by the time of writing it was supposed to be part of the vocabulary of quite many Americans. CBD was uniquely used in the United States until well after the Second World War (Fogelson, 2001).

Wile both Murphy (1972) and Fogelson (2001) use downtown and CBD in a virtually synonymous sense, interestingly, there is a heated debate about the actual use and meaning of the two terms going on a WikiProject called 'Urban Studies and Planning discussion board'[6]. The question is whether or not the downtown and the CBD should be treated in separate articles. The merge is proposed with the argument that the two terms are colloquially identical, with CBD being the Australian and British equivalent of the American downtown. It is strongly opposed by users with a background in urban planning, arguing

[6]http://en.wikipedia.org/wiki/Talk:Central_business_district, accessed 24th October 2008

that the downtown and the CBD are fundamentally different concepts in academic geography. While the first includes the comprehensive civic, cultural and economical functions of the classical city centre, the latter covers solely commercial functions. The two being spatially congruent in some cities does not imply that they are the same thing. There is also some debate whether the concept of the downtown fits European city centres and if CBD is actually used for British cities, as suggested by some contributors. It is hoped that some issues about the terminology for the geographic city centre will be clarified within the scope of this thesis.

2.3 Web-based tagging systems

User-generated content and user-supplied textual labels to categorise online content, a novel approach of organising information, have become increasingly popular on the web. By now, there can be no doubt that social software and tagging systems in particular are not just a fad. While Flickr, for instance, had 375'000 users as of May 2005 (Weinberger, 2007), there are more than seven million of them at the time of writing. This increasing popularity has at length stimulated debate about benefits and weaknesses of tagging in the blogging community and has recently also been discussed in academic research (Macrgegor and McCulloch, 2006).

2.3.1 Definition and characteristics

'Tagging' designates the process by which many individual resource users assign sets of freely chosen keywords or category names to online content such as bookmarks, images, or videos. It is a standard feature of numerous web services like Del.icio.us[7] for bookmarks, CiteULike[8] for bibliographies, Last.fm[9] for music tracks and Technocrati[10] for weblogs, just to mention some of them. Such systems can basically be described by a model of interrelated *resources*, resource *users* and annotated *tags* (Marlow et al., 2006). Platforms supporting tag suggestion based on previously added keywords are referred to as collaborative tagging (Golder and Huberman, 2005) or suggestive tagging systems, as opposed to blind tagging systems (Marlow et al., 2006).

The motivation behind tagging is generally thought to be twofold; apart from the storage and organisation of content for personal means, users of tagging systems are driven by the idea of social contribution and the desire to share with family or the wide public. The two motivations are not exclusive, but the majority of tags are created in a social context, aimed at enabling others to discover and navigate the contributor's resources (Ames and Naaman, 2007). Still, the intention of sharing is not always altruistically motivated; many

[7] http://delicious.com/
[8] http://www.citeulike.org
[9] http://www.last.fm
[10] http://www.technocrati.com

(semi-)professional photographers, for instance, take Flickr as an effective and inexpensive promotion platform for their pictures. The desire to attract the attention of as many people as possible may lead to 'tag spamming' the annotation of a huge amount of prominent but in the context senseless tags. Nonetheless, investigations have shown that most users take tagging seriously (Ames and Naaman, 2007).

Tags act as metadata for personal recall and public discovery, but unlike hierarchical taxonomies or professional indexing, the vocabulary system in tagging is entirely flat (Mathes, 2004). It is also different from the idea of the 'semantic web', the limitedly successful attempt to design formal metadata to organise online content (Steels, 2006). Tagging lacks any control by system administrators, meaning that the user is completely free in adding text strings that seem applicable to the content being marked. As tagging is typically distributed and uncontrolled, the textual labels are susceptible to redundancy, futility and low quality. Shortcomings as found by Guy and Tonkin (2006) in about 40% of Flickr tags include meaningless concatenations (tags are usually processed as single strings and letter case is ignored by the database), misspellings, and highly personalised keywords making sense to single or a small group of users only. Mathes (2004) has suggested that tags follow a power law distribution at the global level of a collection, meaning that there are few tags used by many and a huge number of tags just employed by individual users. The convergence to a small number of frequently employed tags by a community has lead to the creation of the neologism 'folksonomy' from 'folk taxonomies' by Thomas Vander Wal in a mailing discussion list in 2004 (Smith, 2004). Folksonomies are particularly likely to consolidate in suggestive tagging systems (Marlow et al., 2006). The appropriateness of the expression is debated, as tagging systems are missing the typical structure and hierarchy of taxonomies (Golder and Huberman, 2005). Instead, the tagging communities themselves have set up loose tagging rules many users try to follow[11].

The review of literature related to the tagging phenomenon reveals an ongoing discussion about its usefulness and effectiveness. Polysemy (words with many related senses), synonymy (different words with the same meaning), inconsistent usage (e.g. television versus tv) and to a lesser extent homonymy (words with several different meanings), are major problems for the effective information retrieval (Golder and Huberman, 2005). For some authors, as e.g. Shirky (2005), there are literally no synonyms in tagging systems as users are thought to employ distinct expression for very specific and unique reasons. The process of associating tags by choosing the most suitable labels to an item with a potentially uncertain category is a question of making sense and giving meaning to the represented reality. As a matter of fact, it is closely related to the basic level problem discussed in Section 2.1.3 (Golder and Huberman, 2005). The low cognitive cost and effort involved in tagging has played an important role in the proliferation of the tools (Mathes, 2004). At the same time, the choice of tags influenced by possible derivations in the individual constitutions of the basic level is one of the reasons tagging does not perform well with respect to recall and precision (Golder and Huberman, 2005). Tagging systems are not primarily laid out for effective search and retrieval but rather for explorative navigation

[11] http://flickr.com/groups/central/discuss/2026/, accessed 13th September 2008, contains community-generated suggestions of best practices to annotate photographs.

promoting serendipity, the potential of unexpected discovery (Sturtz, 2004). Also Guy and Tonkin (2006: 2, 7) point out that the strength of uncontrolled tagging is that "items can be categorised with any word that defines a relationship between the online resource and a concept in the user's mind" and the "ability of any given user to describe the world as he or she sees it". Textual tags are a direct manifestation of the conceptual and linguistic structure of the user community and its diverse geographical and cultural background (Guy and Tonkin, 2006). It is the ability of tagging systems to capture people's vernacular and view of the world that is particularly relevant and interesting in the context of this project. Due to the populace participating in tagging systems, the databases have been designated as a 'powerful manifestation of distributed knowledge' (Steels, 2006) or as the 'emergence of collective intelligence' (Weiss, 2005). It is assumed that given the existence of a sufficient number of contributors, the system reflects the 'wisdom of the crowd' and that the underlying reality can be derived from consensus (Weiss, 2005). Even though these statements lack scientific verification, tagging systems doubtlessly provide a rich source of empirical data and a novel challenge to the research of cognitive structures and distributed cognition (Steels, 2006).

2.3.2 Geotagging

Relating information to location is an important component of our daily lives (Hill, 2006). Geocoding of hypermedia, for instance, makes web documents searchable and locatable in geographic space and was intensely discussed in conjunction with the location awareness of the semantic web, but is still lacking widely accepted standards. A special type of user-contributed metadata, referred to as *geotagging*, denominates the assignment of spatial references to objects in digital collections. As everything else in the tagging world, geotags are not only dependent on technical configurations but primarily on the user, in particular his willingness and ability to create and utilise geographic information and maps (Erle et al., 2005).

A popular example of geotagging is the spatial referencing of photographs, an obvious operation as imagery is by nature tied to location, at least in terms of a capture position. As the location of the image content can be ill-defined, georeferencing of pictures is by convention a reference to the capture position and not to the image location (Erle et al., 2005). The Flickr community has from the beginning added informal georeferences by means of place annotations, which have been identified as a major category of tags (Winget, 2006). A popular convention was the encoding of coordinates in machine tag mode in the style of geo:lat=43.67736, geo:lon=-79.63236. At present photos can be formally tagged with exact latitude and longitude by using tracklogs from an external GPS, cameras and phones with built-in GPS (Erle et al., 2005), through the Flickr application programming interface (API), or by manually placing the photos on a map interface. Flickr assigns an accuracy level ranging from 1 (world level) to 16 (street level) to every geotagged picture. While the automated approach yields, depending on the GPS signal, a quite precise georeference, the geotag level of the manual approach depends on the zoom level

applied when uploading the picture on the map[12].

2.3.3 Previous work with Flickr

A large number of qualitative and quantitative analysis of tagging behaviour has been carried out by means of the photo-sharing platform Flickr. In the attempt to develop an approach to visualise tag evolution with time, Dubinko et al. (2006) found that the Flickr community had, at that time, on average created more than one million tags per week. According to Winget (2006), these tags consist of all parts of speech which, in turn, typically fall into one of the following categories:

1. Date and time
2. Geographical
3. Narrative (traditional catalogue keywords like building, urban, city)
4. Characterisations (of people or of situations)
5. Individually defined tags (typically compound tags as unique markers)

It has generally been assumed that about half the tags appear only once within a tagging database. In contrast, Guy and Tonkin (2006) found that only 10-15% of the tags were unique in a sample data set from Flickr, a fact they ascribe to the constant growth of the database. A preliminary analysis by Wood et al. (pers. comm.) included all tagged items posted within the bounding box of the UK at highest geotag resolution. On average, every user had uploaded 46.1 items, while the median was only 6, meaning that 73% of the resources in the Flickr sample were submitted by only 10% of the users. 237 out of the topmost 1'000 employed tags within the sample were identified as toponyms. An evaluation of contributor ubiquity for the same tags ranks london[13] first, while england respectively uk are at rank 6 and rank 11. Innercity is at rank 934. While previous work evaluating Flickr's performance of precision in terms of tag-based queries has reported a very low value of 50% (Kennedy et al., 2006), Winget (2006) considered, depending on the query, 80-97% of the 100 most-interesting results as being relevant. Winget (2006) has also performed reliability checks for user-employed place tags against the TGN, a thesaurus which lists preferred and alternate place names in a hierarchical manner. The analysis of tags related to images showing volcanoes revealed that most of the tags corresponded to the preferred name of the respective volcano while the majority of annotations also included alternate names from the TGN. The degree to which Flickr users included the hierarchical structure of place name descriptions was also measured, finding that nearly all the geographic terms ranging from the continent to the volcano name from the TGN hierarchy occurred in the Flickr sample. Schmitz (2006) presented a preliminary method to induce ontology from

[12] http://www.flickr.com

[13] Whenever referring to a tag from Flickr, the respective word will be displayed in sans-serif text style.

Flickr tags based on a probabilistic subsumption model. The method deriving parent-child relations for tags was successfully tested for place annotations.

Of particular interest are the investigations taking advantage of the spatial component of formal geotags from the Flickr database. Early results by Girardin and Blat (2007) show that the type of city, respectively urban landscape, possibly influences the granularity applied upon manual geotagging. Some frequently tagged cities have their peaks of granularity apparently at the city and the street level, while for others there is no predominant location resolution. Familiarity with a place, on the other hand, does not seem to have any impact on the geotag level. In another experiment, Girardin et al. (2008) discriminated Flickr users into locals and visitors to compare their digital 'traces' left by means of georeferenced photos within the city of Rome. The authors were able to designate the main locations of tourist activity and compared the spatial distribution of tag semantics, for instance of ruins, to the actual cityscape. A series of related projects performed by research associates at Yahoo![14], the company currently owning Flickr, were aimed at generating meaning from spatial tag patterns. These approaches are completely data driven and not dependent on gazetteers, predefined lists of landmarks or a manual classification of tags. Rattenbury et al. (2007) present a technique for the automated determination of tags corresponding to events or places. Tags with specific spatial patterns representing locations at different scales were successfully extracted by the *burst detection method* from *Spatial Scan statistics* or the specially developed *Scale-structure Identification*. The approach was extended by Kennedy et al. (2007) who present a location-tag and vision-based technique to generate a set of images which are representative for previously identified place tags. Ahern et al. (2007) automatically determine representative tags for geographic regions by a k-Means clustering algorithm. Candidate terms are subsequently scored by term frequency (the number of times a tag was used within the cluster) – inverse document frequency (the overall ratio of a tag amongst the items within the entire region under consideration). Highly ranked keywords, typically corresponding to places or landmarks, are displayed on a tag map, referred to as aggregated 'psychological map' by the authors. This technique was integrated into the 'World Explorer'[15], an online map application representing tags according to their relative prominence at the respective location.

2.4 Conclusion and research questions

The review of literature has shown, that in order to design effective geographic information technologies and enhance GIR, we need to know how people perceive, think of and describe space in an intuitive way. Attention within GIScience and GIR has been drawn to the automated understanding of place names and to the implications of such multiple and vaguely defined places with regard to their satisfactory modelling and formalisation. It has also been revealed, that the production, explanation and modelling of intrinsically vague geographic regions is not only a concern since the emergence of the user-centred

[14] http://research.yahoo.com/
[15] http://tagmaps.research.yahoo.com/worldexplorer.php

perspective in GIS, but has long occupied scholars of various branches of geography. With respect to urban space, the internal structure of spatially differentiated cities, resulting in more or less homogeneous, but typically ill-defined sub-regions, has occupied quantitative, behavioural, and cultural geography, cognitive and spatial sciences, as well as GIS and information science.

Various disciplines have developed their own ways to gain insight into the characteristics, configurations, and extents of urban districts and the city centre in particular. Functional geography adopted quantitative techniques to delimitate the CBD by means of statistical data or objective, measurable categories. The intention was to partition urban space into clearly-defined, functional districts allowing for universal comparison, regardless of local conceptualisations and names of the districts (Murphy, 1972). Cognitive and behavioural studies revealed, that despite the inherent fuzziness of cognitive categorisation as well as of geographic entities, there is broad agreement between the individual understandings of places, their names and their extents. Many authors have accentuated the importance of such vernacular places as shared frames of reference of a culture and society (Rapoport, 1976; Talen, 1999; Hill, 2006; Davies et al., 2008; Sen, 2008). Motivated by research questions of GIScience and information retrieval, Montello et al. (2003) and Mansbridge (2005) have adopted behavioural approaches to get insight into people's understanding of vernacular regions of the city, namely downtown Santa Barbara and the city centre of Sheffield, respectively.

The most easily observable artefact of human conceptualisations of geographic space is natural language. The regional geographer Byrkit (1992), for instance, took advantage of textual descriptions in publications to explore the manifold definitions of the American Southwest. Due to the prevalence of spatial data on the Internet, the web itself is currently the major source of information about the constitution of vague regions. The aim is to support the resolution of user-employed place names in the context of information services and GIR by modelling the places by means of vague or crisp footprints. Most of the authors (Purves et al., 2005; Arampatzis et al., 2006; Jones et al., 2008; Henrich and Lüdecke, 2008) mined the web as a whole to derive knowledge of large-scale geographic regions. Schockaert and Cock (2007) as well as Twaroch et al. (2008) used hints about places at the neighbourhood level in Yahoo! local data, and georeferenced business directories, respectively. Finally, Grothe and Schaab (2008) derived footprints for vague regions such as the Rocky Mountains and the Alps from georeferenced photos posted on Flickr.

Given the nature of geographic entities and human spatial knowledge, it was stated in different contexts that spatial terminology is dominated by ill-defined referents without precise semantic as well as geographic boundaries (Hirtle, 2003; Hill, 2006; Purves and Jones, 2007; Davies et al., 2008). The main question is how people, depending on their socio-linguistic background and given circumstances, categorise and name places and how this behaviour can be imitated by computer systems. Previous studies have focussed on the human conceptualisation of natural entities (McGranaghan, 1990), or the place names used within web queries. These indications were analysed with respect to frequency (Sanderson and Kohler, 2004), granularity (Zhang et al., 2006), the query length, and the kind of subjects they were related to (Gan et al., 2008). An investigation by Davies et al. (2008)

with end users of map products revealed that people usually mean urban locations when speaking about places. Neighbourhoods and districts are obviously considered as the most important referents of place. To date, it has not been established on a large-scale how people actually describe place at the sub-city level of granularity.

Therefore, within the scope of this project the Flickr database is taken as a information source of how humans employ place names and how they understand ill-defined places. The creation of ad-hoc keywords to categorise georeferenced content on Flickr is seen as a proxy of how people intuitively refer to location. Flickr can in some respect be seen as a case study of tagging systems in general, which were characterised as direct manifestations of conceptual and linguistic structures of the user community (Guy and Tonkin, 2006). Even though it is possible to mine any text associated with photographs on Flickr, this study is restricted to the analysis of tags in order to gain insight into tagging as a categorisation process and on the capabilities of tagging in the context of vernacular geography. In this context, the lack of pre-defined categories and restrictions is considered as beneficial, as it allows users to describe location exactly in the way they intuitively categorise a place. That users contributing to the empirical data are not aware of being participants of a study ensures maximum possible intuition in the employment of tags. Hence, the first research question is formulated as follows.

How do people describe urban places in terms of tags depending on language and culture region?

As the worldwide database provides a unique option to investigate common patterns as well as intercultural differences, the first (as well as the third) research question will primarily be addressed by means of the urban core, which was established as an important component of the city structure in virtually all urbanised cultures (Hofmeister, 1996). In the previous literature review across geographic fields city centres have been identified as "almost archetypal examples of geographic objects with indeterminate boundaries" (Thurstain-Godwin and Unwin, 2000: 2), regardless of the viewpoint and the approach adopted.

Due to the chaotic and uncontrolled nature of tagging, the reliability of the user-generated metadata has been controversial in other contexts. The information challenge facing tagging systems is to extract knowledge from unstructured sets of tags. Despite the lack of ontology and semantics, Rattenbury et al. (2007) showed that the patterns in Flickr metadata allow for the automated extraction of tags corresponding to place descriptions. Grothe and Schaab (2008) successfully modelled vague footprints of large-scale geographic regions by means of density surfaces. These experiments imply that the users' employment of tags emerges in a spatial and textual structure which is somehow consistent and correlated. The statement by Steels (2006: 287), who considers tagging systems "as a tremendously powerful way to coordinate the ontologies and views of a large number of individuals, thus constituting the most successful tool for distributed cognition so far" remains speculative and cannot be considered as thoroughly verified. The initial findings of Wood et al. (pers. comm.) suggest that we have to expect considerable bias through single contributors in the metadata pattern. The aim, addressed by the second research question, is to explore

the spatial aspect of formal and informal georeferences of Flickr tags in more depth and to establish the conditions necessary to derive a collective cognitive view of vernacular regions.

Is user-generated metadata in online photo-collections suitable to capture vernacular geography?

To date, it has not been investigated if users' attitudes towards the creation of metadata are sufficient at the neighbourhood level of granularity which were identified as the most important types of places for end users of map products (Davies et al., 2008). The contribution of this thesis is to establish what kind of textual place descriptions occur in terms of tags and whether the tags and the formal georeferences are accurate enough to derive knowledge of places at the sub-city level of granularity. This is particularly questionable due to the essential nature of photography; people are likely to take and tag pictures of environments they are not particularly familiar with. The second research question includes an investigation of the quality and errors in geotags, which has not been thoroughly evaluated to date.

Footprints are a powerful means for cataloguing and retrieving spatially related information (Larson, 1995; Hill, 2006; Purves and Jones, 2007). The mapping of vague and vernacular regions close to human cognition in order to populate gazetteers used for information retrieval, information systems, and map services is currently the major motivation for the study of people's beliefs about vernacular places. Consequently, the third research question has been formulated as follows.

How can digital footprints of vernacular regions be modelled from georeferenced tags?

Compared to 'ordinary' web resources, the main advantage of the Flickr data are the direct links between location and textual descriptions of place. The costly geoparsing and geocoding process can be skipped. It is also basically different from the empirical tests performed by Montello et al. (2003) who asked people to sketch neighbourhood boundaries on a map. As pointed out by the authors themselves, considerable bias might be introduced to the results by the map section chosen for the experiment. Furthermore, the review of spatial cognition literature revealed that environmental knowledge is typically topological in nature, but fuzzy in terms of boundaries. There are hints (Ferrari, 1996; Hill, 2006) that we rather have the declarative knowledge necessary to assign a location to a specific point in space than being able to delimit a region's boundary on a map. Basically, the nature of the Flickr data used in this project corresponds to the conception of the study by Mansbridge (2005), but the immense amount of data available on Flickr allows for scalability, in terms of considered places as well as number of participants. It will be part of the project to explore to what extent the users are influenced by the map displayed upon posting images.

As established in the background chapter, this work is situated between vernacular conceptualisation of space, which is by nature inaccurate, vague, and incomplete and computational modelling and formalisation in GIS and GIR. It has been stated on various occasions that it depends on the context, the purpose, and the user of the model how boundaries are constituted and should be represented (Burrough and Frank, 1996; Davies et al., 2008). Therefore, a mode of representation which is adapted to the geographic objects under consideration, the underlying data, and the purpose of the model had to be established in the course of this project. The vagueness inherent in the Flickr data is mostly of the contested form. The fact that some parts of space will be considered as typical representatives for a region by many users, while other places will only be tagged with the place name by few, is represented in the density of the point pattern. Another aspect introducing uncertainty are possible outliers, caused by erroneous data or by tags associated with pictures showing the place from far away. The challenge was to derive suitable footprints from point-wise information representing the emergence of collective cognition in the spatial context.

Regarding the evaluation of the technique of footprint approximation, we face a problem inherent to all representation of vague geographic entities, which can never be mapped and discretised with certainty (Couclelis, 1996). As shown in the background sections, cues about the nature, location, and dimension of a city's subregions are derived from a variety of factors, such as the built structure, land use, social homogeneity, population density, and housing systems. The interplay and influence of these many factors explains that residents' designations of neighbourhoods may differ fundamentally from administrative definitions, which are not physically visible, but will still affect the residents (Campari, 1996). Other boundaries, marked by rivers or major roads, are clearly manifested in urban space (Lynch, 1960; Smith and Varzi, 2000). The benchmark for the approach of footprint approximation is therefore performed by means of public parks. Parks are spacious objects within the city environment which, unlike city neighbourhoods, can be considered as spatially well-defined regardless the mode of observation and regionalisation. Also the evaluation of the underlying data, described in the following chapter, will be accomplished accordingly.

Chapter 3

Data

3.1 Flickr

Except as noted otherwise, the information in this sections was directly taken from the various webpages under the Flickr domain[1]. The basic functionality of Flickr has originally been developed for a multiplayer web-based game, named 'Game Neverending'. In 2004, the photo-sharing and tagging applications were incorporated into Flickr which has gained increasing popularity over its short lifespan. The service was purchased by Yahoo![2] in 2005 and is today the most popular photo sharing application and community on the web (Winget, 2006).

3.1.1 Design and characteristics

On Flickr not only the tags, but also the resources are user-contributed. Each resource, corresponding to a digital photograph, video, or image in the wider sense, has a range of settings and controls associated with it. Images are uploaded to user-specific *photostreams* through the website, a mobile phone, or the API and may be organised in *photosets*, corresponding to the traditional photo album. When uploading pictures, the interface prompts the user to add descriptive features such as a title, a caption, and tags. Flickr belongs to the category of blind tagging systems (Winget, 2006). The number of tags per item was earlier on restricted to 75 but has subsequently been deregulated (Wood et al., pers. comm.). By default, only the user owning an item has tagging rights. Even though others might be enabled to annotate ones own items, a negligible subset of tags are not created by the resource owners themselves (Marlow et al., 2006). Other metadata typically associated with Flickr items are the capture and upload time, the owning user, comments, a usage license, and for geotagged items coordinates as well as a place indication (Figure 3.1).

[1] http://www.flickr.com/
[2] http://info.yahoo.com/center/us/yahoo/

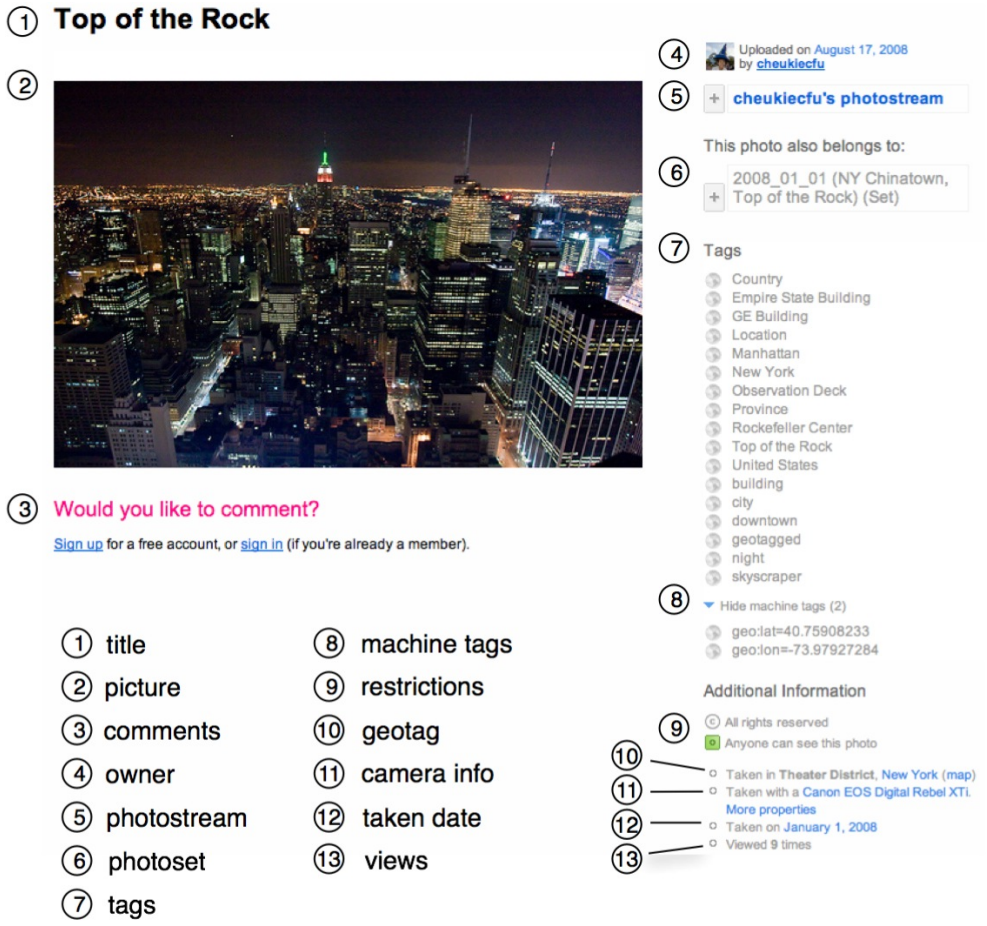


Figure 3.1: Example of metadata associated with georeferenced photo on a Flickr page (Source: http://www.flickr.com/photos/cheukiecfu/2769751529/)

Each user's homepage links to the respective photosets, a user-specific world map, and a profile page, indicating, if specified, personal information such as special interests, profession, or the place of residence. Flickr allows users to specify a family, friend or contact distinction and to define restriction levels for photos, determining which resource can be accessed by whom. According to Weiss (2005), less than 20% of the images uploaded to Flickr are restricted from public viewing. The privacy level for geotags can be set independently from the photo privacy enabling the user to hide the photo location from strangers.

The tagging utility is a major focus of the system. It allows for browsing, navigation, and exploration of ones own photostreams, specific user's collections, the entire database or of a specific place. There are 'popular-pages' where the most frequent tags at the global level, by individual users, or within a geographic place are depicted as 'tag clouds'. For more focussed search, a text-based search returning items matching a query and a specific tag-search function are implemented. The results may be listed by 'most recent' or 'most interesting', a quality that is defined by a secret algorithm. Flickr provides a cluster function, based on a related-tag algorithm to support the disambiguation of homonymous

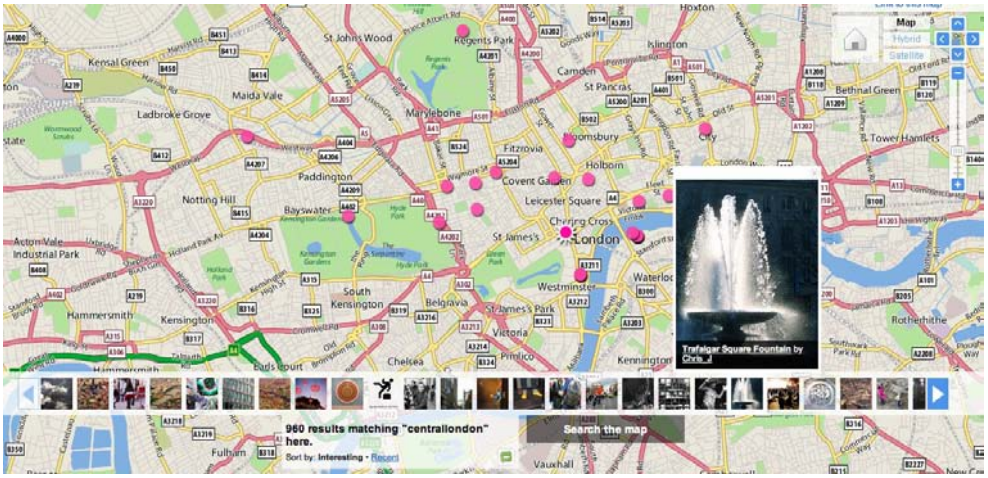


Figure 3.2: The Flickr map interface with a selection of georeferenced items (Source: http://www.flickr.com/map/)

tags. According to Flickr, the related-tag feature is based on clustered usage analysis and appears to be more sophisticated than a technique of simple term counts.

The requirements of the many users adding machine tags in the style of geo:lat=43.67736, geo:lon=-79.63236 for location information were satisfied when the developer team incorporated standard geotagging facilities into the platform on August 28, 2006. Already by the next day, 1.6 million georeferenced images were uploaded onto the platform (Winget, 2006). To date, almost ninety million geotagged photos have been posted on Flickr[3]. The geotagging process, as explained in Section 2.3.2, is supported by the *world map*[4]. The *places*[5] application integrates all items posted to a specific area of the map. Flickr places relies on hierarchical gazetteers whose resolutions vary substantially between different parts of the world. The gazetteer is based on a set of overlapping bounding boxes and special algorithms are employed in order to decide on the named location corresponding to the geotag. This process is referred to as *reverse geotagging* by the Flickr engineers. Due to the vague nature of many geographic places, particularly at the neighbourhood level, a feature was recently incorporated allowing users to adjust the place designations made by the database. The aim is to adopt people's local knowledge by integrating users' suggestions into the spatial database in order to enhance reverse geotagging (Catt, 2008).

Spatial filtering of Flickr items is possible by means of the interactive map (Figure 3.2), by using a theme and named place as a search specification on the map, or by the 'exploration' of places through the respective application.

[3]http://code.flickr.com/blog/2008/10/30/the-shape-of-alpha/, accessed 14th November 2008
[4]http://www.flickr.com/map/
[5]http://www.flickr.com/places/

3.1.2 Data collection from Flickr

All functions and metadata described in Section 3.1.1 are made available through the API published by Flickr[6]. Well-documented interfaces have stimulated vast development activity in the community yielding a wide range of applications based on Flickr services and data[7]. Within this project, the API was employed in order to access an amplitude of empirical data that exceed the capacity of manual browsing.

City-dataset

city	download start[a]	items found	retrieved	% retrieved	tags/item
Zurich	May 02 10:34:01	47'162	47'005	99.7	5.4
London	May 24 02:07:54	1'080'036	1'061'883	98.3	5.4
Sheffield	May 18 01:32:44	21'127	21'124	99.9	5.6
Chicago	May 23 11:00:35	392'659	389'703	99.2	4.9
Seattle	May 16 12:12:14	313'796	313'796	99.4	4.9
Sydney	May 04 21:33:28	140'387	139'542	99.4	5.2
grand average					**5.2**

[a]Central European Standard Time

Table 3.1: Collection of georeferenced items within the bounding boxes of the different cities in the city-dataset

The Flickr API supports a search function that can be adjusted by the specification of a range of parameters. Unauthenticated calls of the method will return metadata in XML-format associated with publicly available photos matching the search criteria. In order to access the REST-based[8] data returned by the Flickr database, the Java wrapper provided by the Flickrj API[9] was employed. Flickrj is open source software and can be downloaded for free from the sourceforge.net[10] platform. At the time of implementation, the place-based methods were 'experimental' in the Flickr API and not yet included in the Java wrapper classes. Instead, bounding box restrictions were used to filter the search by location, indicating the bottom-left and the top-right corner by coordinates represented in the WGS84 system, the reference system used in Flickr for storing the point locations of geo-tagged items. Initially, the Flickr database was sampled for all publicly accessible images in the bounding boxes of selected cities without any restriction on tag content, subsequently referred to as city-dataset (shown in Table 3.1). The coordinates used for the investigated cities were manually extracted from Google Earth[11] and are indicated in Appendix A.1. The Java program was then adopted to mine the database for georeferenced items matching a tag restriction, used for the filtering of generic city core tags on the entire globe

[6]http://www.flickr.com/services/api/
[7]http://www.flickr.com/services/
[8]REpresentational State Transfer
[9]http://flickrj.sourceforge.net/api/
[10]http://sourceforge.net/projects/flickrj/
[11]http://earth.google.de/

(global-dataset in Table 3.2). Finally, items associated with a specific tag, respectively city toponym from a certain culture region, were collected to get an extended sample for the different regions (region-dataset in Table 3.3). Also, Java scripts were implemented to access the related-tag method and to extract the home location for a list of users.

Global-dataset

generic tag	download start	items found	retrieved	% retrieved	tags/item
downtown	Jun 02 07:31:00	133'885	133'655	99.8	11.8
central	Jun 02 22:53:44	29'067	29'042	99.9	15.5
cbd[a]	Jun 02 17:43:36	4'442	4'422	99.5	15.4
innercity	Jun 02 18:02:29	2'360	2'359	100.0	20.7
citycentre	Jun 03 11:42:29	2'057	2'057	100.0	11.9
citycenter	Jun 03 11:30:38	949	949	100.0	15.7
grand average					**15.2**

[a]including centralbusinessdistrict

Table 3.2: Extraction of items with specific tags georeferenced anywhere in the world (bounding box -180°, -90°, 180°, 90°)

Region-dataset

culture region	download start	items found	retrieved	% retrieved	tags/item
GB cities[a]	Jun 27 10:21:57	5'048'437	4'140'602	82.0	6.0
US cities[b]	Jun 15 11:20:27	1'986'434	1'682'648	84.7	6.4
AUS cities[c]	Jun 04 14:13:28	918'667	859'176	93.5	8.6
grand average					**7.0**

[a]birmingham, liverpool, glasgow, edinburgh, london
[b]chicago, seattle, boston, miami, houston
[c]brisbane, sydney, perth, melbourne, adelaide

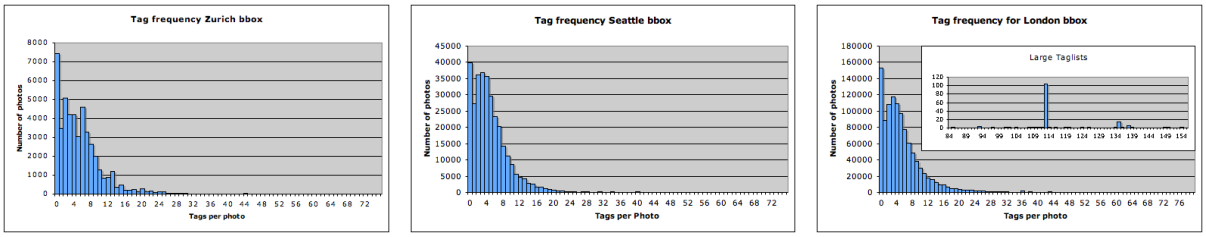
Table 3.3: Collection of items matching any of the cities in the respective list of city toponym tags

A problem well known to the API group is the restriction that the Flickr servers will keep returning identical hits if more than 5'000 items match a query. To avoid the problem, the results were requested by decreasing upload date and the time span in the search criteria was adapted automatically in a way that not more than 1'000 were extracted at once. Repeating the query until the entire time span was covered did not yield all photos originally matching the criteria, as the instances are imprecisely sorted in the Flickr database. The more iterations required to extract a dataset, the lower was the recall achieved. The numbers of retrieved items in Table 3.1 to Table 3.3 give an idea of the sample size finally available for the different evaluations. For retrieved matches, the associated tags in the 'clean' version as processed by Flickr for the construction of URLs and additional

metadata was extracted. The snapshots were stored in text files with each line representing a collected instance and the columns representing the metadata required for evaluation.

3.1.3 Data properties

Quantitative analysis of the extracted data was also accomplished by means of Java programming and the resulting text files were imported into Microsoft Excel for visualisation. As shown in Figure 3.3 for the bounding boxes of Zurich, Seattle, and London, the number of tags associated with each photo converges towards a bi-modal distribution, the larger the underlying data sample. On a global level, there seems to be a peak of many photos with zero tags assigned and a second skew population with a median of about three tags. The bulk of pictures has less than twenty tags associated, but the larger the sample, the higher is the likeliness for tag spamming to occur. The frequencies of tags per item for the global- and the region-dataset, which obviously yield at least one tag per item, are shown in Appendix A.2.1. While the very large samples in the region-dataset exhibit a regular skew distribution, the picture for georeferenced tags with specific city core terms of the global-dataset is heavily disrupted. Possible explanations therefor are discussed in Chapter 5 and 6.



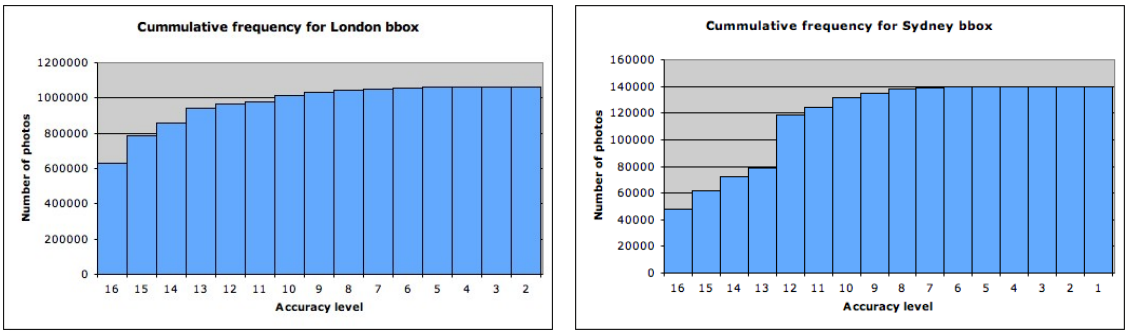
(a) Zurich, N = 47'005 (b) Seattle, N = 313'796 (c) London, N = 1'061'883

Figure 3.3: Frequency of tags per item for georeferenced data sets in the bounding boxes of different cities

bbox	total tags	different tags	unique tags
Zurich	253'768	14'046	56.5%
London	5'782'340	187'196	46.1%
Sheffield	119'085	10'319	53.1%
Chicago	1'910'862	68'659	46.9%
Seattle	1'523'550	60'293	46.9%
Sydney	722'151	30'575	47.0%
average			**49.4%**

Table 3.4: Tag statistics for georeferenced items within the bounding boxes of different cities

A grand average of 5.2 tags per item with a very low variance was found for the city-dataset in Table 3.1. The global-dataset exhibits a higher variation of comparatively large values yielding a mean of 15.2 tags per instance (Table 3.2). 7.0 tags were identified on average for the region-dataset (Table 3.3), yielding an overall mean of 9.1 tags per item. Thus, the trend can be said to agree with the 7.1 tags per photo calculated by Wood et al. (pers. comm.), who rejected the resources with zero tags for analysis. For the city-dataset, also the proportion of unique tags was calculated, which is displayed in Table 3.4. There seems to be a slight tendency that the proportion of tags appearing only once decreases for larger data sets. But overall, almost 50% of the tags were determined as unique. The numbers in Table 3.4 imply that the growth of the Flickr database does not substantially increase consistency.

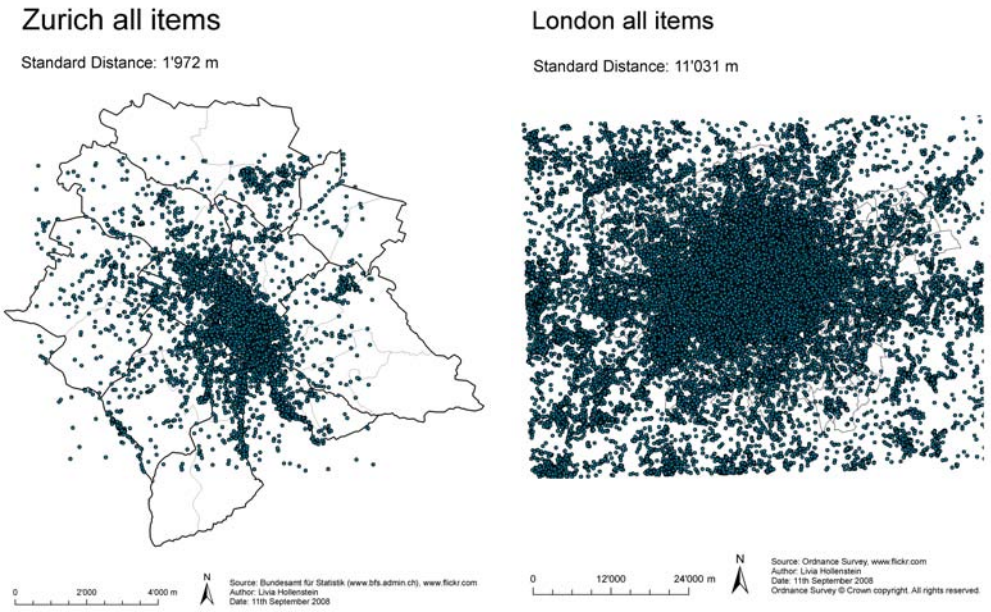


(a) London bbox (b) Sydney bbox

Figure 3.4: Cumulative frequencies of geotag accuracy in the bounding boxes of different cities

The georeferenced city- and global-dataset were also investigated with respect to the applied geotagging resolution. The diagram in Figure 3.4(a), reflecting the cumulative frequency of geotag accuracy levels found in large samples, shows that the vast majority of items is posted at a resolution between street accuracy (level 16) and city accuracy (level 12) with decreasing frequency towards lower granularities. This is also generally true for the other snapshots of different cities in the city-dataset (Figure A.4 in Appendix A.2.2). An exception is the Sydney sample in Figure 3.4(b) which has a peak at level 12, while the more accurate geotag levels are underrepresented. This observation is in accordance with Girardin and Blat (2007) who found that different cities exhibit specific distributions of location resolution applied by the users. In our case, the variation is simply explained by a difference in the level of detail in the backdrop mapping and the unavailability of zoom levels higher than 12 for the maps in the Sydney area at the time of data mining, which will be discussed in more detail in Section 5.2.2. To summarise, the initial analysis of the accuracy level specified by users in the geotagging process seems to suffice the purpose of this project.

Visual display of all items found within the bounding boxes of the city-dataset reveals that the majority of geotagged images are located in the vicinity of the geographic and tourist core of the respective cities (Figure 3.5 and A.6 in the Appendix). The concentration towards the centre is more pronounced for cities with smaller data bases. The bias has



(a) Zurich (b) London

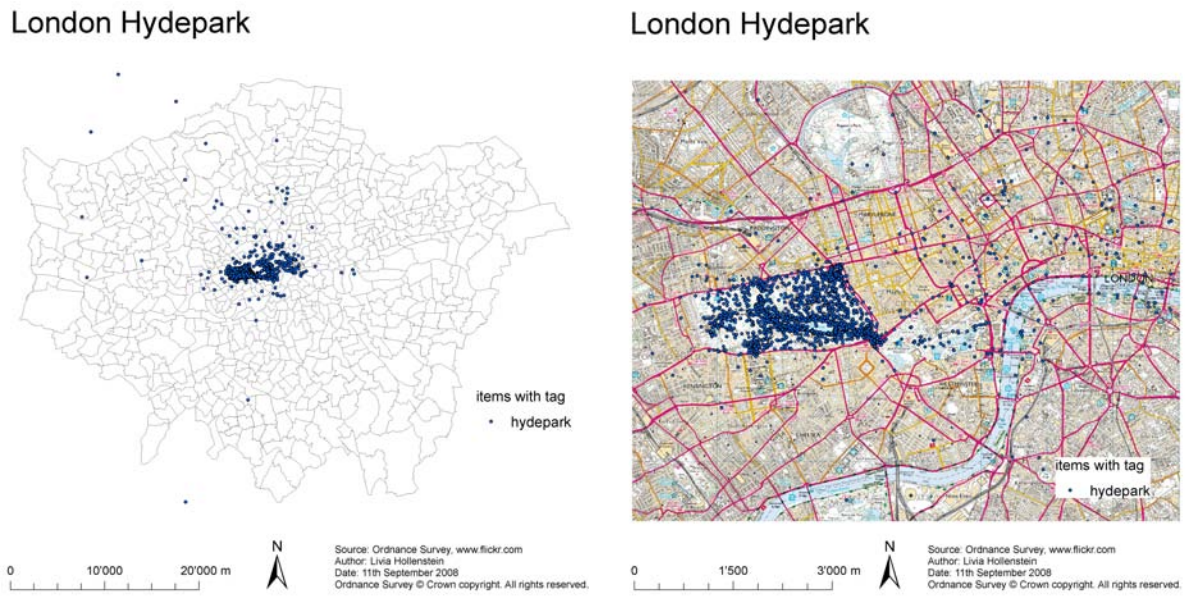
Figure 3.5: Spatial distribution of all georeferenced items within the bounding boxes of Zurich and London

also been observed by Girardin and Blat (2007) for the city of Barcelona and has to be accounted for in further analysis of the spatial distribution of tags.

3.1.4 Evaluation of user-generated content

The review of previous work based on Flickr revealed contradictory predications with respect to the quality of user-generated content. Therefore, an evaluation of the data collected from Flickr with a special focus on the quality of the geotags was carried out at this stage. Due to the limitations of human observation and the subjective structuring of space in the process of spatial cognition, we have to expect substantial variation in the employment of geotagged annotations. As discussed previously, this applies also to human conceptions of theoretically well-defined, official regions. Hence, suitable for an evaluation of user-generated data is a place delimited by a boundary of the more bona fide sort, provoking minimal disagreement on its location and extent.

Therefore, Hyde Park in London was chosen to examine the reliability of georeferenced Flickr data. For this purpose, all items tagged hydepark within the bounding box of London were extracted, yielding a set of 9'775 instances. At first sight, the hydepark data in Figure 3.6(a) appears to be randomly distributed within the central part of London. On closer examination, the points exhibit substantial auto-correlation and overall match the shape of Hyde Park surprisingly well, while exhibiting local clusters near the Speakers' Corner in the southeast and around the Serpentine, as shown in Figure 3.6(b). Obviously, a lot



(a) (b)

Figure 3.6: Distribution of georeferenced items tagged hydepark *in London*

of users do not distinguish between Hyde Park and the Kensington Gardens, the western part of the park, which has been technically separate from Hyde Park since 1728[12]. As the two are not divided by a bona fide boundary, the whole area is considered to belong to Hyde Park in the following evaluation.

8'484 items of the data set were posted within Hyde Park as a whole or no further than on the main roads surrounding the park, displayed red in the backdrop map. 1'291 or 13.2% of items are presumably misplaced. The correctly placed instances were uploaded at accuracy levels between 10 and 16 with an average of 14.15 and a standard deviation of 1.4. The geotag granularity of the outliers range from 3 to 16 with a mean of 11.13 and standard deviation of 3.5. As expected, this suggests that reliable geotags are generally employed at higher resolution. As evident in Figure 3.6, misplaced items are generally biased towards the geographical centre of London and are frequently placed in nearby Green Park or St. James's Park. Apparently, less attentive users either put the items near the centroid of the London map or are unable to identify the proper park, probably due to having an inappropriate map section or zoom level displayed. Actually, 42% of the outliers stem from only three contributors which have posted 196, 184, and 172 items, respectively. One user has obviously automated the geotagging process, as all 196 pictures from his London photoset are placed in the correct location, but have the identical taglist assigned. The users contributing 172 and 184 photos have simply placed their pictures in the wrong park, namely in Hampstead Heath and St. James's Park. Their photos are responsible for the spurious peaks in the density surface of the data represented in Figure 3.7. Evidently, a few prolific and careless users significantly distort the picture.

[12]http://www.royalparks.org.uk/parks/kensington_gardens/, accessed 9th July 2008

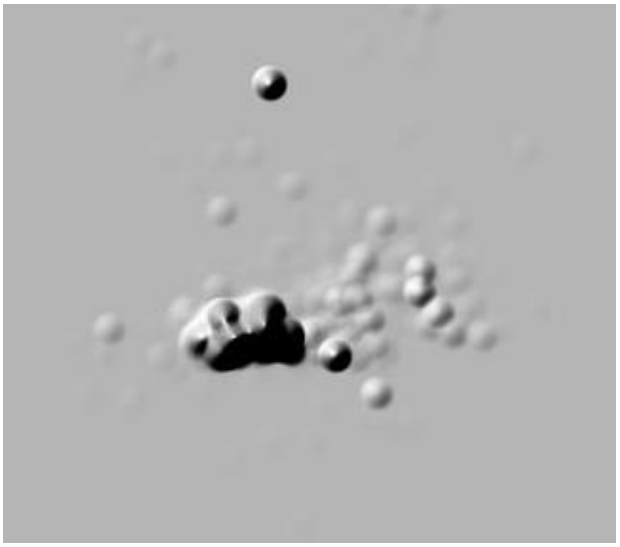


Figure 3.7: Hillshade representation of hydepark *data within London*

In order to analyse the reasons for inaccurate data in more depth, also with respect to place tags in natural language, 100 random outliers posted by 100 different users were analysed by looking them up on the Flickr website. The experiment revealed three major reasons for the anomaly between location and tag:

1. The capture location is not identical with the location information in the tag. In the case of Hyde Park, **7** outlying pictures were taken from airplanes and did amongst others show the park. This case cannot be considered as low quality data.

2. The photos are tagged correctly but are misplaced on the map. This was the case for **69** items, whereof **17** were placed in another park.

3. The tag choice is apparently incoherent. **22** items tagged hydepark did not have an obvious relation to the park.

In the third case, users have mostly added the same taglist to a whole set of photos uploaded at once. Except for the three scenarios described above, there was one photo collage including many different places within London and another one for which the tag choice made sense in the context, as the owner had obviously been to a concert at Hyde Park that night. While context-dependency was an issue, an ambiguity problem did not occur in the case of Hyde Park within London. This second analysis revealed that about 7% of the outliers are justified, yielding that approximately 12% of the total georeferenced items relating to Hyde Park within London must be suspected for misplacement. The relation of misplaced and mistagged items suggests that most users obviously do take tagging seriously but not all of them are willing or able to correctly locate on the map. The evaluation confirms that the overall attitude of the users towards the creation of metadata complies with the aims of this project. But the fact that very little users might produce a lot of and possibly erroneous data has to be considered carefully when investigating

the nature and location of user-employed tags believed to reflect common knowledge and cognition.

3.2 Other data

Additional data was used as a backdrop for visualisation and verification of the data extracted from Flickr. For the comparison at the global level, polygon layer shapefiles representing the boundaries of the continents created by ESRI Data & Maps[13] were used. The world maps are referenced in the WGS84 system and were projected into the cylindrical equal-area Behrmann projection with ArcGIS 9.2. The polygon shapefile for Zurich originally stems from the statistical office of the Canton of Zurich. The area of the municipality is composed by twelve administrative 'Kreise' (districts) which in turn contain two to four official 'Quartiere' (neighbourhoods). The backdrop data for the United States (US) was retrieved from Zillow[14], a company providing web-based evaluation of real estate. Zillow has created boundary data for nearly 7'000 neighbourhoods of the largest US cities by integrating information from various institutions such as local authorities, online sources, and real estate companies. Just like Flickr, Zillow is interested in integrating users' suggestions for boundary improvement. The boundaries can be downloaded as shapefiles under a Creative Commons License and were then transformed into the equal-area Albers USG projected coordinate system. The backdrop data for England was provided by Ordnance Survey, referenced in the British National Grid. It includes the official boundary lines of the Westminster Constituencies as well as 1:25'000 scale colour raster data, showing detailed environmental features and annotation. A selection of four, and six 10 km by 10 km tiles were provided for Central London, respectively Sheffield. As the tiles for London do not cover the entire area of Greater London, the backdrop mapping for the different footprints are inconsistent. Comparative data was used for visual inspection of the generated surface models on various occasions and will be introduced at a later stage.

[13] http://www.esri.com/data/data-maps/overview.html
[14] http://www.zillow.com/webtools/labs/neighborhood-boundaries.htm

Chapter 4

Methodology

4.1 Analysis of place tag usage

The examination of quantitative aspects of tag employment is aimed at gaining insight into the means and expression used to describe urban space in terms of tags. To verify the theoretical assumptions about cultural and linguistic variation with respect to the vague urban core, visualisations of data sets associated with six selected English terms which denominate the city core was performed at a global level. In order to investigate the colloquial usage of vague spatial concepts and vernacular place names in more detail, six cities from different language and culture regions were chosen as a basis for quantitative evaluation of place tags. The city of Zurich, a representative of the German language area, was chosen for an in-depth analysis of tags due to familiarity reasons. London, Sheffield, Chicago, Seattle, and Sydney were selected as representatives of the British, Anglo-American, and Australian culture regions. Apart from cultural aspects, the selection was motivated by the intention to find out how the choice of tags, respectively concepts, is influenced by the nature and size of the urban environment, as well as to thoroughly evaluate Flickr for the purpose of capturing vague terminology under different circumstances. Quantitative analysis was achieved by means of Java programming and the results were illustrated using Microsoft Excel.

4.1.1 Tag profiles of contribution ubiquity

The evaluation of the reliability of user-contributed data in Chapter 3.1.4 has shown that bulk uploads are a common occurrence on Flickr and that a few users may introduce major distortion into the metadata pattern. Therefore, a method "to explore the possible effects of bias by prolific or unprolific posters" was adopted from Wood et al. (pers. comm.) to evaluate the commonness of specific tags at the global and at the city level. The technique is based on the construction of a 'tag profile', as shown in Figure 4.1. To construct the tag profiles, the instances within a data set are sorted according to their owners' rate of

contribution, with the most prolific posters on the left and the less active contributors on the right side of the x-axis. The items are binned into groups corresponding to one-hundredth of the total number of instances in the data sample. Hence, several bins on the very left might be contributed by the same user, while the rightmost column is made up of items owned by many different users. Subsequently, the occurrence of the tag in question is counted for every bin and represented by the size of the blue bar.

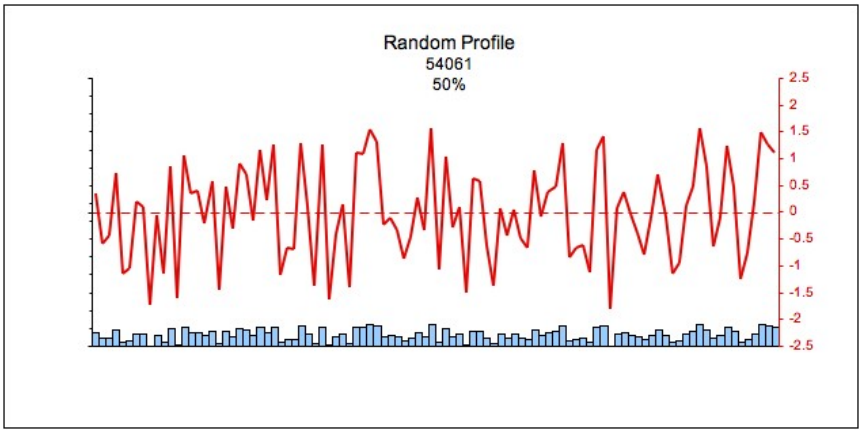


Figure 4.1: Profile for a tag with a random distribution. Blue bins represent absolute occurrences of the tag, the red line is the z-score normalisation. The upper number indicates the number of occurrences of the given tag within the sample, the lower number corresponds to the coefficient of variation.

In order to compare the different patterns of contribution, standard or z-score values are computed to normalise the counts of the tag in a bin according to $z = \frac{x-\mu}{\sigma}$, where x is the raw count of the respective tag in the bin, μ is the mean of the tag per bin, and σ is the standard deviation of the entire population with respect to the given tag. The z-score is represented by the red line with peaks corresponding to bins that contain a higher-than-average proportion of instances with the given tag, and with lows corresponding to a representation of the tag below the average.

The overall bias in the employment of tags by different user groups is expressed by the coefficient of variation, defined as the ratio of the standard deviation σ to the mean μ of the population. The coefficient of variation serves as a measure of whether tags are employed with equal frequency among users of varying prolificness. It rises with decreasing ubiquity, meaning that tags, whether frequent or not, only used by a small group of people will show high coefficients of variation. In this sense, the technique is applied to analyse if a term is widely used in order to fulfil the requirements of capturing an aggregated view rather than the perspective of individuals.

4.1.2 Tag clouds of co-occurrence

The analysis of tag usage to describe the urban core at a global level was accomplished by the investigation of tag co-occurrence. For this purpose, the total number of instances

of each distinct tag was calculated in the different samples of the global-dataset. For all snapshots, each associated with a specific city core term, the thirty most frequent tags corresponding to toponyms were identified and checked against the geonames gazetteer[1]. As a listing of thirty tags and numbers did not seem meaningful, the usefulness of tag clouds as a means of information visualisation was tested by representing the frequencies of the place tags in the global-dataset this way. A standard online service[2] developed by a doctoral student in Design and Education at Stanford University was used to generate the tag clouds of co-occurrence from a text file which reflected the pattern of place tags in a data snapshot.

Tag clouds have emerged along with tagging systems and are widely used tools in web applications supporting user-generated metadata. The clouds serve as a visual model of tag usage and are laid out as interfaces for the exploration and navigation of large data sets. In a tag-cloud, keywords are usually listed in alphabetical order and displayed at a size and style corresponding to the relative prominence of the respective tag within the data set (Hassan-Montero and Herrero-Solana, 2006). An example of a tag cloud, not

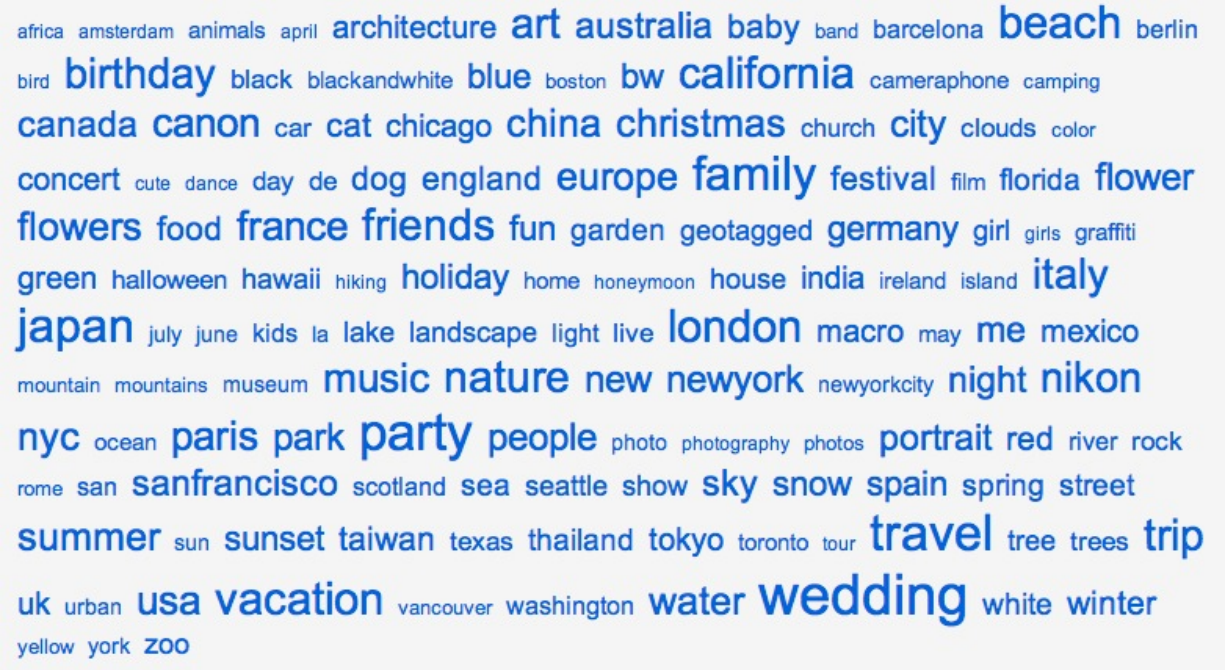


Figure 4.2: Tag cloud representing the all time most popular tags on Flickr, extracted on the 6th of July 2008 (Source: http://www.flickr.com/)

representing the frequency of co-occurrent tags, but the absolute popularity of tags, is the all time most popular tag page of Flickr. The illustration in Figure 4.2 shows that tagging images with place names is indeed popular. As of July 2008, 44 out of the 144 of all time most popular tags referenced to proper toponyms.

[1] http://www.geonames.org/
[2] www.tagcrowd.com

4.1.3 Analysis of frequency counts

In order to determine whether the counts of distinct terms used to describe the urban core vary significantly between different language areas in the region-dataset, a χ^2-test for homogeneity was employed. Additionally, the connection between vague terminology and the provenance of the users, as well as the correlation between the natural language tag and the geotag granularity was tested this way.

The χ^2-test for homogeneity is used to determine if the frequency counts of a variable (in the columns) are distributed equally over different populations (in the rows) of a contingency table. For every column, there is a null hypothesis stating that the proportion of the variable is the same in each population. The alternative hypothesis says that at least one of the null hypotheses is false. The test may be applied if the variable is categorical and the expected frequency for every cell of the contingency table is at least five. Expected counts for the cells are calculated and the test value is derived from the sum of difference between observed and expected frequencies. If the test value is higher than the critical value at the chosen level of significance with the given degrees of freedom, the null hypothesis can be rejected, meaning that the counts are not equally distributed. The test statistics of the χ^2-test of homogeneity were calculated in Microsoft Excel according to the formulae in the AP Statistics Tutorial[3].

4.1.4 Identification of place tags

An analysis over the entire area and time span of the bounding boxes was achieved by producing a list of tags and their occurrences in descending order of frequency for each of the six cities in the city-dataset. To enable a manual classification method of tags, only the 1'000 top-ranked tags were classified in the samples of the Anglo-Saxon cities. The tags occurring within the bounding box of Zurich were analysed completely in order to obtain a full count for quantitative comparison. The tags were first distinguished into place tags and non-place tags and the latter categorised according to the granularity level represented in the place concept. Given the unconfined possibilities to encode place information in terms of unstructured tags, some conventions about the constitution of a place tag had to be made. For the purpose of this study and due to the nature of the data, a distinction between geographic terms as defined by Sanderson and Kohler (2004) and actual place names was made according to the following rules.

- Names representing places which belong to the hierarchy of continent-state-county-city (additionally districts and streets for Zurich) and descriptive terms of location such as city, citycentre, and neighbourhood were regarded as place tags. Toponyms not related to the city under consideration (e.g. vienna in the bounding box of Zurich) and indications in the form of precise coordinates (e.g. geo:lat=47.3722) were not counted as place tags.

[3]http://stattrek.com/AP-Statistics-4/Homogeneity.aspx?Tutorial=AP, accessed 18th September 2008

- Landmarks and geographic features such as parks, lakes, airports, and locative adjectives like australian, british were not considered as places.
- Accordingly, names of institutions, buildings and events, e.g. zürifescht and universityofchicago were not regarded as place tags.
- Following the procedure of Gan et al. (2008), interpretable misspellings were included and counted with the correctly spelt version of the place, because the user's intention when assigning a tag is considered to be relevant for the purpose of the evaluation, which is not intended as an analysis of the correctness of user contributed-tags.
- Different languages and compound expressions, e.g. zurich, zürich and zurich2007, chicagoatnight were included and the compound tags counted with the instances of the corresponding simple place tag.
- Compound tags were allocated to all granularity levels represented in the tag, e.g. bahnhofstrassezurich was considered to belong to the street and the city level, as apparently for the user it was important to express that is not any 'Bahnhofstrasse' and at the same time not any location in Zurich.

In case of doubt, candidate tags were checked against the geonames gazetteer, the web service map.search.ch[4], as well as additional Internet resources. The identification of city core terminology was guided by the theoretical background, but all tags were considered as possible candidates in order to take advantage of serendipity. In this way, each tag was labelled according to it representing a continent, a country/state, a city, or a generic city core concept. The relative frequency of occurrence in relation to the total sum of tags for Zurich (the 1'000 top-ranked tags for the Anglo-Saxon cities, respectively) was calculated for every place tag and summed up over the correspondent level of granularity.

In order to get an analysis of tag statistics not only at the global level of the bounding boxes, but also in terms of place tags added per single item, every instance within a bounding box was checked whether having *at least one* of the established place tags at a specific granularity in its taglist.

4.2 Analysis of spatial tag distribution

4.2.1 The standard distance

The standard distance tool provided by ArcGIS can be used to compare the compactness of different distributions by a single value representing the degree to which points are concentrated or dispersed around the mean centre by means of a distance[5]. As the bulk of georeferenced Flickr data is strongly concentrated at the geographic centre of the cities, the

[4]http://map.search.ch/
[5]http://webhelp.esri.com/arcgisdesktop/9.2/

standard distance was employed to verify whether the instances associated with a specific term are more concentrated towards the centre than the total of the georeferenced instances in the bounding box.

Figure 4.3: Illustration of the standard distance (Source: http://webhelp.esri.com/arcgis desktop/9.3/, accessed 15th of November 2008)

4.2.2 Visualisation of vague footprints with KDE

From the methods suggested in Section 2.1.6, kernel density estimation (KDE) was adopted to model the geographic extent of vernacular places from georeferenced tags. The field-based model represents uncertainty by continuously varying values, reflecting the gradation of the approximated reality. If required, thresholding of the surface may yield a single or a set of sharp boundaries. KDE is a standard feature in many GIS programs. To process the Flickr point data, it was imported as text files into ESRI ArcGIS 9.2 and the WGS84 coordinates were transformed into the respective local, metric reference systems.

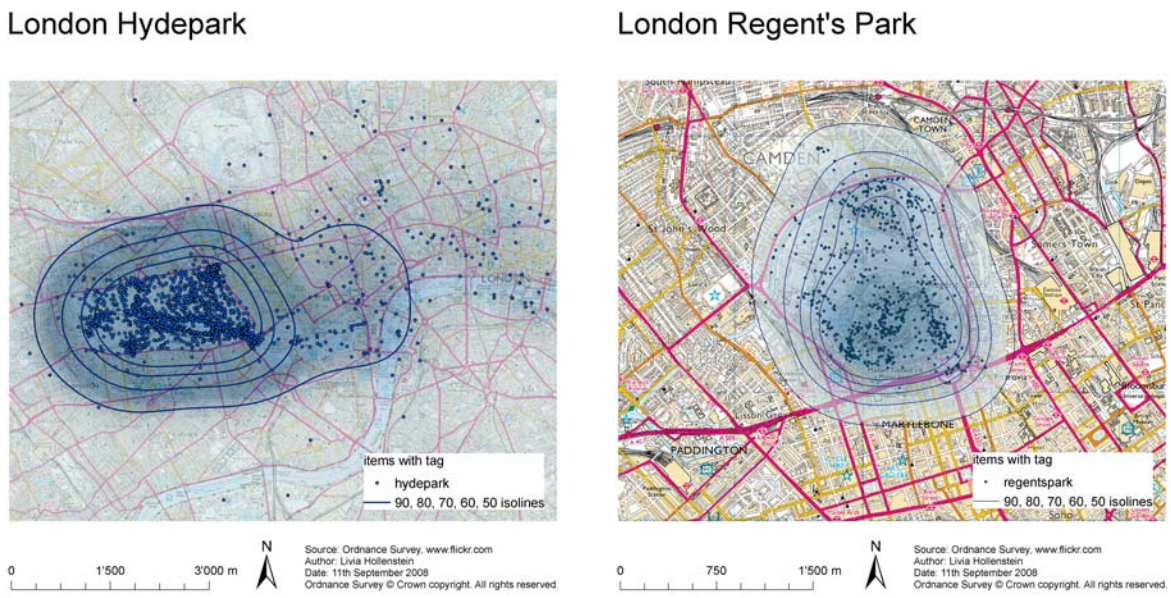
Kernel density estimation methods are used to produce field representations of local density estimates from two-dimensional point distributions. The density value is estimated at each observed point by spreading the search radius by some type of kernel function with defined bandwidth. The kernel function weights the points within the search radius as a function of their distance from the kernel centroid. By adding the values from each kernel at a point over a finely drawn grid, a useful visual indication of estimated densities is obtained. The method requires the selection of a range of parameters that influence the resulting surface. While the kernel function and the grid size do not have a major impact, the choice of the kernel bandwidth, also termed smoothing parameter (de Smith et al., 2008), and in our case the choice of the threshold value to exclude possible outliers, will strongly affect the

result. A large bandwidth tends to 'oversmooth' and extend the pattern, while a small radius will result in strongly focussed and disjoint surfaces (de Smith et al., 2008).

Actually, the choice of the bandwidth turned out to be a critical aspect, as there is no agreement on how to approach this problem in classical GIS literature. While O'Sullivan and Unwin (2003: 87) state that "generally, experimentation is required" and de Smith et al. (2008: NN) that "Bandwidth selection is often more of an art than a science", the treatment of possibly 'false' data, as in the case of the outlier problem in the Flickr data, is not addressed at all. Previous efforts to approximate vague regions by density estimators treated the problem of the kernel parameters in various ways. As the studies were mostly concerned with single regions, many authors have experimentally determined the bandwidth to a size that represents the underlying point pattern best (Purves et al., 2005; Jones et al., 2008), or use the same search radius for all regions within a city environment (Thurstain-Godwin and Unwin, 2000; Twaroch et al., 2008). Also the threshold point density to generate sharp boundaries is usually defined interactively. Jones et al. (2008) start with an initial threshold of one point per grid cell and progressively half the value until a single-peaked surface remains. The guidelines benchmarked by means of official regions are applied to map the extent of vague places. A data-driven approach was presented by Henrich and Lüdecke (2008). An investigation of 39 well-defined regions revealed that the optimal threshold correlates strongly with the maximum value of the density function and is subsequently determined automatically by a linear function. Grothe and Schaab (2008) determined a heuristic for the expected fraction of outliers by Support Vector Machines (SVM) by analysing the results for well-defined boundaries in the form of eleven European countries.

Due to the large number of data sets to be processed within this project and due to the uncertainty about the nature of the underlying concepts, an experimental choice of kernel parameters was not considered appropriate. Early attempts made clear that the bandwidth had to be determined from the underlying point pattern. A solution was finally found in the field of wildlife research, where kernel density estimators are used for GIS-based analysis of animal home ranges. The concern for an objective choice of the smoothing parameter has motivated the development of applications providing a data driven determination of the kernel bandwidth, as for instance the Home Range Tools (HRT)[6] designed for ArcGIS. HRT supports fixed and adaptive kernel estimators based on the Gaussian kernel function and provides several automated and objective methods for defining a suitable bandwidth (Rodgers and Carr, 1998). For the reasons explained in Section 3.1.4, again, parks in London were taken for the validation of the approach. Additionally, the technique was tested by the vaguely defined region of the centrallondon sample. To minimise the effects of both bulk uploads by single users, internal clusters, and erroneous data as described in Section 3.1.4, all x- and y-multiples and all items geotagged at an accuracy level lower than nine were removed from the data sets in a pre-processing step. The fixed kernel methods were found to produce better results, as adaptive kernels tend to assess undesirable local clusters. Due to computational overhead, the least sophisticated of the HRT methods to automatically define the search radius was applied. It takes a standard distribution h_{ref}

[6] http://blue.lakeheadu.ca/hre/



(a) (b)

Figure 4.4: Footprints for Hyde Park and Regent's Park with 90%, 80%, 70%, 60%, and 50% isolines and geotagged items from Flickr

as a reference for the bandwidth parameter, which is calculated from the mean variance in the x- and y-coordinates as follows.

$$h_{ref} = n^{-1/6}\sqrt{\frac{var_x + var_y}{2}}$$

The h_{ref} method is suitable if the underlying point pattern is unimodal, i.e. single-peaked (Rodgers and Carr, 1998) which should usually be the case for georeferenced tags referring to places at the sub-city level. Visual inspection revealed that by this method the search radius is, despite the outliers, chosen in a way that appropriately represents the extent of the well-defined parks by visual comparison (Figure 4.4). For the public parks as well as for the widespread pattern of centrallondon, the bandwidth was determined to a value almost identical to the radius that was established experimentally. As the resolution of the raster is typically chosen to be smaller than the bandwidth parameter, it did not significantly change the characteristics of the calculated surfaces. It was set dependently to the size of the total area under consideration (i.e. 10 meters for Zurich and Sheffield and 50 meters for London, Chicago, and Seattle). Also, the classification method imposed on the surface values did not have considerable impact. Finally, a standard deviation classification with an interval of 1/3 standard deviations was chosen for mapping.

The outlier problem was addressed in an objective way by means of the volume contour feature provided by the HRT. The contour lines connect points of equal density whereas the outermost line surrounds an area which belongs to the region under investigation with a probability of 0.9. Figure 4.4 illustrates that the 50% contour line most closely approximates Hyde Park while the 90% line produces a considerable overestimation. For

Regent's Park the 50% volume contour is too narrow as a restriction. Based on these two examples it becomes obvious that, depending on the point pattern and the shape of the object, different volume contours delineate the represented objects best. The exact shape of the features is never obtained, due to the heavily biased point distribution within the parks. While the references for Hyde Park are biased towards the Speaker's Corner in the southeast, the point pattern for Regent's Park is clustered towards Madame Tussaud's in the south.

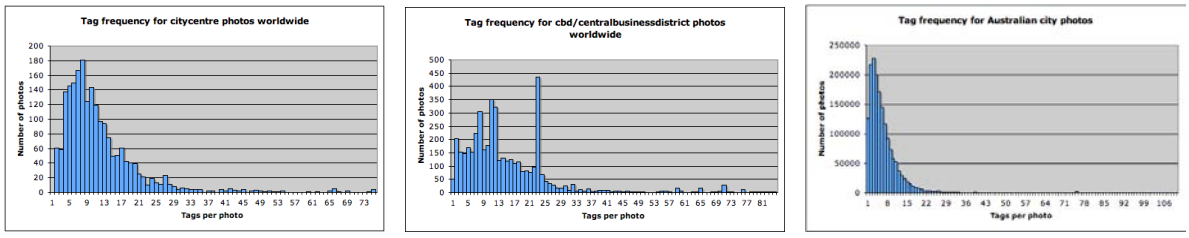
The surfaces representing ill-defined regions were cut off at the 90% contour line, as illustrated by the footprint of Regent's Park in Figure 4.4(b). The 90% isoline will most certainly contain outliers and tend to overestimate the regions. Hence, the derived models do *not* correspond to exact footprints that could be used in gazetteers, but rather to a means of representing all facets of uncertainty, imprecision, and differing conceptualisations. The procedure is aimed at providing a basis for an objective comparison and discussion. The evaluation of vague regions is challenging as they cannot be compared to a single, valid boundary. It is accomplished through manual assessment of results by comparing to previous attempts of delineating, describing, or defining the region under investigation. Suggestions for further expansion of the approach towards automatic generation of footprints will be made in Sections 5.3 and 6.3.

Chapter 5

Results and interpretation

5.1 City core terms at the global level

Before analysing the tags used to describe the vague urban core in the Anglo-Saxon culture regions, some comments are made on the metadata properties of the items in the global-dataset, which were filtered by means of different city core tags as displayed in Table 3.2 in the Chapter Data. An evaluation of user bias by means of tag profiles is not possible for this data set, as every instance in the sample contains the tag under consideration. However, it was found that the most prolific 10% of users contributed 57% of the citycentre data, and 70% of the cbd data. For innercity, they even account for a proportion of 83%, while one user alone has uploaded 61% of the items. This tendency was also observed in an initial analysis of Wood et al. (pers. comm.). Apparently, it is not particularly related to this data set, but needs to be accounted for in further analysis.



(a) citycentre (b) cbd (c) Australian cities

Figure 5.1: Pattern of tags per item for data sets related to city core terms (a) and (b), and for items related to five specific Australian city toponyms (c)

More striking is the pattern which the samples of the global-dataset exhibit in the tag frequency diagrams as shown in Figure 5.1. While there is a high proportion of items uniquely tagged with the specific toponym used for filtering in the region-dataset, practically all photos associated with a city core term have more than this one tag assigned. As

evident from Figure 5.1(a) and 5.1(b) compared to Figure 5.1(c)[1], the diagram peaks of the global-dataset are generally shifted towards the right of the x-axis, yielding an average of 15.2 tags per item compared to 7.2 and 5.9 for the reference data in the city-dataset and the region-dataset, respectively. Obviously, people rarely describe the location of online resources just by generic city core terminology. These kind of tags are likely being applied in the context of relatively long taglists.

5.1.1 Visualisation of worldwide tag distribution

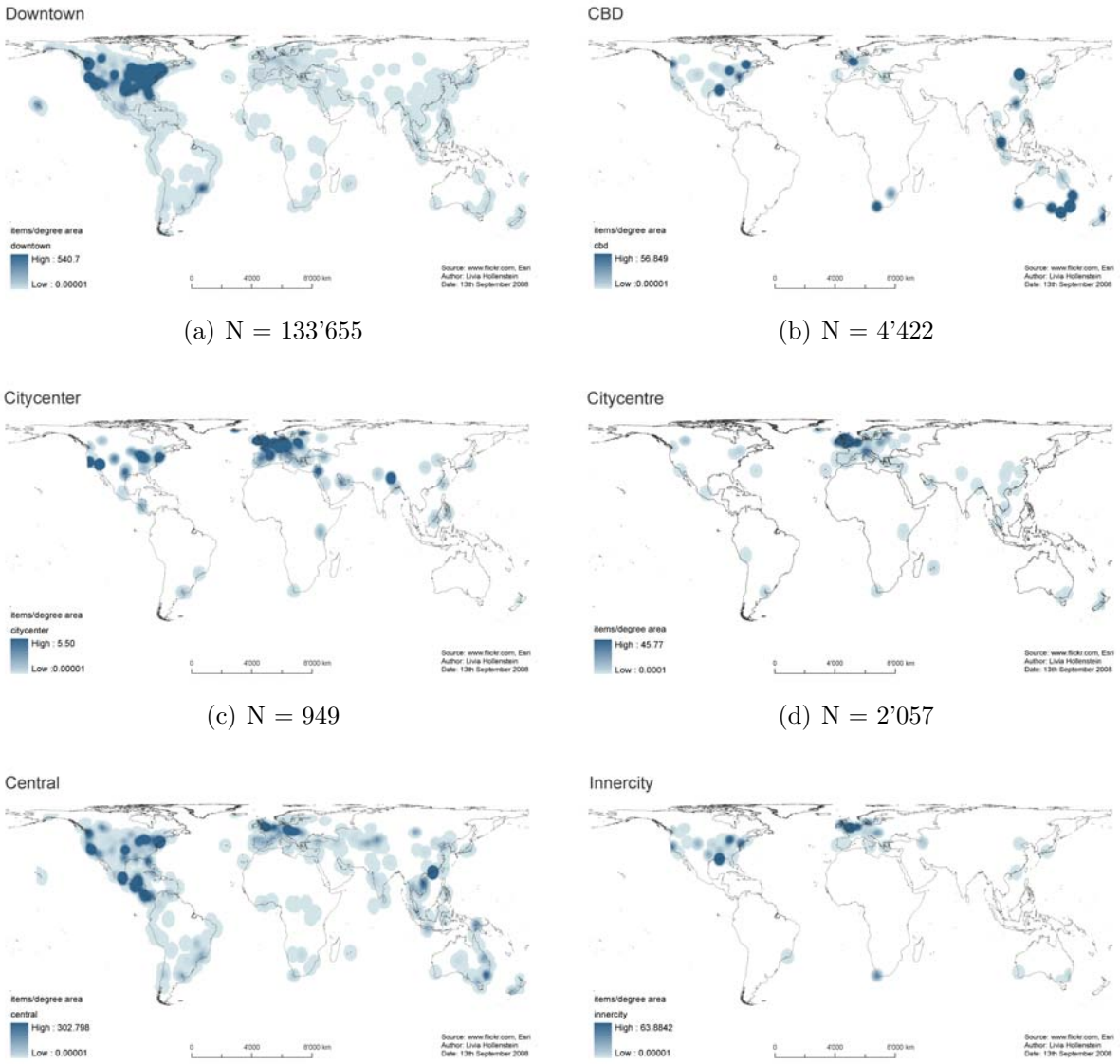
The world maps in Figure 5.2 aim to provide a general impression and qualitative comparison of spatial patterns of terminology usage. The world maps are not generated by data-driven kernel estimators but the search radius was experimentally determined and set to 4.5 degrees for all data sets. Also, multiple occurrences of points at the same location were not removed before estimation. Even though the estimations are based on a scores of instances, the data is therefore susceptible to the bias induced by single users and the maps must be considered carefully.

As implied by theory, the tag downtown is predominantly used in the urban regions of Northern America, but appears in many of the world's regions, with particular peaks in Honolulu and So Paulo. Cbd, on the other hand, is much less frequent and omnipresent than downtown. The centres of all the major cities in Australia as well as Wellington in New Zealand are referred to as CBD by Flickr users. While CBD is also widely used in Beijing, Singapore and Cape Town, it is only commonly employed to specify the business core of New Orleans in North America, where Central Business District is the name of a specific neighbourhood, corresponding to the usual American downtown[2]. The cluster of CBD over Paris is invoked by a single French person referring to the business district of 'La Défense'. The British spelling of citycentre occurs almost exclusively in the UK. Citycenter on the other hand is mostly used in Europe, and in some of the North-American cities, while a single overseas traveller is responsible for the peak in Asia. The filtering for central does not necessarily yield photos related to the urban core, but the distribution of the snapshot suggests that the tag is primarily used as a geographic indication. The major clusters are connected to Central America and Central Europe, but the term occurs also frequently in New York and other major cities in the US. The remaining peaks represent central parts of London, Sydney, and Hongkong. Two single users are basically responsible for the metadata pattern of innercity, which is overrepresented in London and New Orleans. The activity leading to the second cluster is also reflected in the geotag diagram of the innercity population shown in Figure A.5(c) in the Appendix, which exhibits a major peak at accuracy level 9.

The global pattern of tag usage does generally reflect the theoretical assumptions well. Though, the investigation reveals that some expressions, such as inner-city, are rarely employed in colloquial language.

[1] Further examples are given in Appendix A.2.1

[2] http://gnocdc.org/orleans/1/47/index.html, accessed 15th October 2008



(a) N = 133'655

(b) N = 4'422

(c) N = 949

(d) N = 2'057

(e) N = 29'067

(f) N = 2'359

Figure 5.2: Visualisation of global densities of selected city core terms as represented in georeferenced tags from Flickr

5.1.2 Evaluation of co-occurrence

The tags depicted in Figure 5.3 represent the 30 most frequent place tags occurring within the different samples of the global-dataset, each associated with a generic city core term. In the process of place tag identification in the taglists some improvisation was required as the tags were inconsistent not only with respect to spelling but also as regards term boundaries. Probably due to the not very intuitive user interface, New York, for instance, occurred as newyork as well as new and york. The picture given by the tag clouds in Figure 5.3 generally reflects the patterns inherent in the world maps described above. Four out of the thirty most frequent tags associated with cbd occurred in a character set not interpretable by the

(a) downtown

(b) cbd

(c) citycenter

(d) citycentre

(e) central

(f) innercity

Figure 5.3: Clouds of 30 most frequent place tags associated with generic city core terms worldwide

system, and most likely corresponded to Asian toponyms. Compared to the world map, North America is hardly represented in the CBD tag cloud, implying that the term is not commonly used there. Central is primarily employed in conjunction with Central America or with city in general. Again, New York stands out, a connection which is probably amplified by the city's relation to Central Park. The Eastern Europe cluster in the central tag cloud is contributed by a single user from the Czech Republic referring to Central Europe. The spatial distribution in the world maps for downtown and citycentre is confirmed by the tag clouds of co-occurrence, but the clusters of citycenter in the US is not represented in the tag cloud. The German zentrum would have ranked 31st among the tags co-occurring with citycenter, implying that users with a German background are likely to employ citycenter when annotating resources in a global platform. The co-occurrence of innercity with specific subregions of London is invoked by a single user and therefore again not considered representative.

5.1.3 Analysis of user provenance

In order to avoid distortion introduced by single contributors, the evaluation of users' provenance contributing to the samples in the global-dataset did not consider how often people employed a given tag. Rather, it was verified whether or not a term is part of a user's vocabulary. A list of unique photo owners was generated for each sample of the global-dataset and the users' location of residence extracted from Flickr. Due to the countless options open to indicate the location of residence (e.g. New York, NYC, Big Apple), the locations were analysed manually. The classification was guided by the concept of the culture regions and checked against the geonames gazetteer. As self-tagging is prevalent, this procedure will mostly reflect the provenance of the person who actually employed the tag. On the other hand, the current location of residence does not necessarily reflect the linguistic background of a person.

	downtown	cbd	citycenter	citycentre	innercity
number of users	9'364	473	199	300	104
with unambiguous location	6'181	317	133	204	66
UK	164	27	13	123	13
Continental Europe	743	29	61	53	24
Oceania[a]	51	157	1	4	2
North America	4'856	55	42	16	24
Latin America	217	1	2	0	1
Asia	136	40	13	6	1
Africa	14	8	1	2	1

[a]Includes Australia and New Zealand

Table 5.1: Provenance of users applying different city core terms of the global-dataset

The seemingly uneven distribution of user counts per region in the snapshots associated with different generic city core terms in Table 5.1 was confirmed by the χ^2-test of homogeneity. The calculated test value of 4'224 largely exceeds the table value of 39.25 at the 0.001 level of confidence for 16 degrees of freedom, meaning that the null hypotheses of even distribution within the columns can clearly be rejected. Even though the counts for Asia, Latin America, and Africa were summed up, the expected frequency within this region was small and broke the 'rule of thumb' concerning the restriction of use of the test.

However, the distribution of users per region in Table 5.1 implies that also on the individual level, there are specific patterns of terminology usage for the vague urban core. As expected, citycentre is predominant in the UK, but also used by people from Continental Europe. The American version citycenter occurs, even in absolute terms, more frequently among Continental Europeans than among Americans, implying that the latter do not often relate to the city core as such, while none-native speakers from all over Europe are likely to translate the notion of 'centrum'. Both versions of city centre are entirely unknown in Australia and New Zealand. Inner-city is very rarely employed by native speakers. The high proportion of inner-city among Continental Europeans was particularly caused by German users, possibly translating the expression 'Innenstadt' literally. CBD is the absolutely dominant term among users from Australia and New Zealand. Most of the Americans and all the Africans using CBD originate from New Orleans and Cape Town respectively, a fact which is also represented in the world maps.

Particularly striking is the fact that people from all parts of the world, except in Oceania, most likely employ downtown of the Anglo-Saxon city core terms to describe the central part of a city. The expression occurs in absolute and relative terms more frequently than any other English expression not only in North America but also on Continental Europe, Asia, Africa, and Latin America. The majority of Latin American users were identified as Mexicans, which are possibly influenced by the geographical proximity to the US. Even users indicating a British hometown would rather employ downtown than the traditional citycentre.

5.1.4 Analysis of data from different Anglo-Saxon culture regions

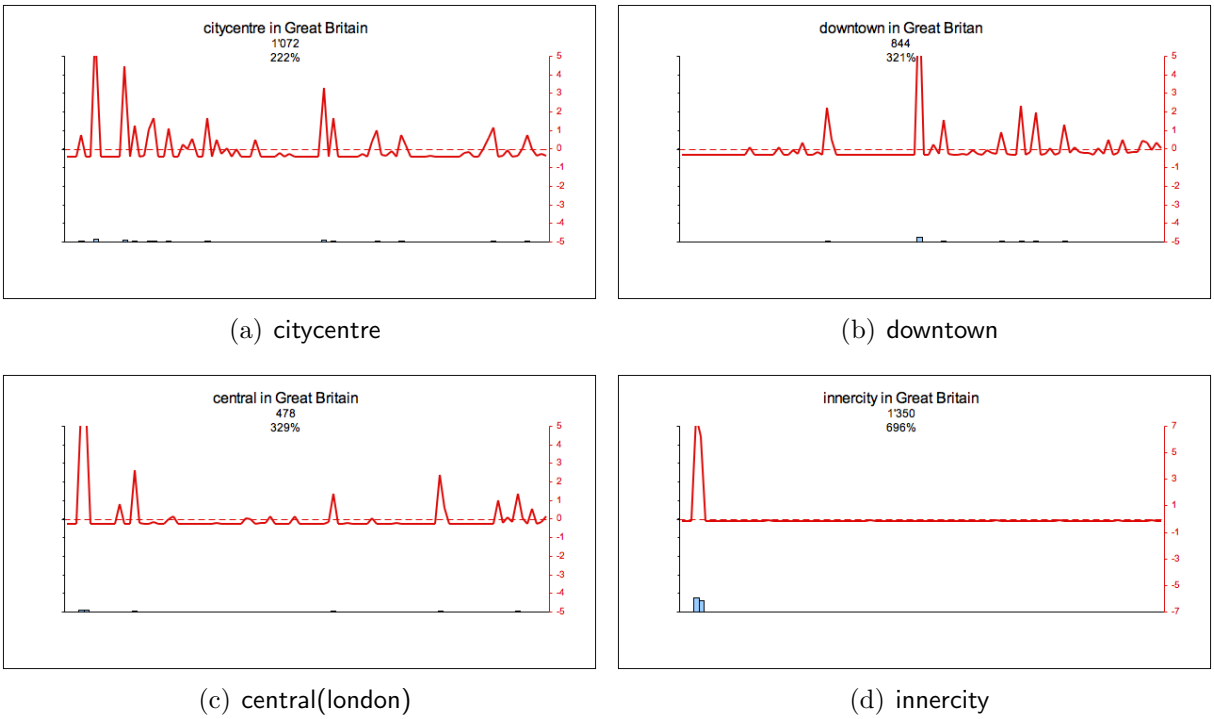
The previous evaluation revealed that the metadata properties of the rather small geotagged sets associated with generic city core tags are partially distorted by single users. Therefore, the database was enlarged by the region-dataset, which was collected by means of tags corresponding to names of major cities in the UK, the US, and Australia (displayed in Table 3.3). A drawback of the region-dataset is that the metadata is to some extent influenced by specific applications of terms in the cities chosen as a filter. The taglists of each of the three samples in the region-dataset were mined for the previously discussed city core tags and all concatenations thereof. A χ^2-test of homogeneity confirmed with a value of 119'450 that the proportion of tag counts in Table 5.2 can be considered as uneven at all levels of confidence.

Within the cities of the UK, citycentre is the most frequently used term, while Downtown

	downtown	cbd	citycenter	citycentre	innercity	central
GB cities[a]	844	4	103	1'072	1'350	478
US cities[b]	88'978	109	194	89	157	0
AUS cities[c]	2'311	10'851	15	130	178	15

[a] birmingham, liverpool, glasgow, edinburgh, london
[b] boston, miami, seattle, chicago, houston
[c] sydney, melbourne, perth, adelaide, brisbane

Table 5.2: Occurrence of city core terms in samples representing cities of different language areas



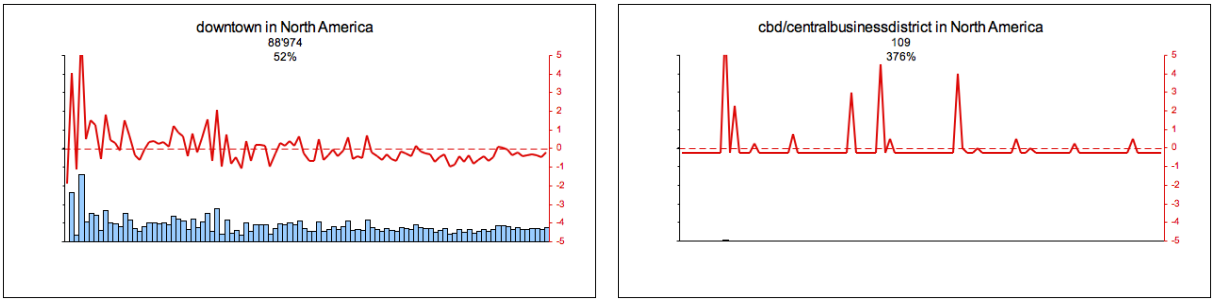
(a) citycentre (b) downtown

(c) central(london) (d) innercity

Figure 5.4: Tag profiles for city core terms associated with different toponyms tags of British cities

occurs nearly as often. Central in conjunction with a city name in the UK is used for London only. Regarding the popularity of the generic tags in the UK, expressed by the coefficient of variation and shown by means of tag profiles in Figure 5.4, the value of downtown (321%) is even marginally lower than the value of centrallondon (329%). Though, the high variation in the employment of downtown among users who posted few photos within the UK might reflect that mostly tourists use this term to describe British city centres. Again, for the frequent term innercity the picture is distorted by a single user (Figure 5.4(d)). As found in the previous evaluations, CBD is definitely not used within the UK, as it occurs only four times in the sample. As expected, in North America (Figure 5.5) downtown is the absolutely prevalent term with a minimal coefficient of variation of 52%. Cbd has a rather

high coefficient of variation of 375% and was less frequently employed than citycenter and even innercity. This is another indication that CBD is not commonly used for the average city in the US. Despite the fact that citycenter only appears 194 times within the American data set, it exhibits a low bias in user ubiquity of 116%. Central in conjunction with a toponym does not occur once in the US snapshot. Apparently, the expression is prevalent in New York only. In Australian cities[3], cbd (101%) is confirmed to be the dominant term. Also downtown is popularly used (209%) and even innercity (259%) is quite common in Australia. Altogether, the counts in Table 5.2 show that the base for evaluation was not essentially extended by this experiment, giving another hint, that except for downtown, and CBD in Australia, generic expressions are rarely employed in terms of tags.



(a) downtown (b) cbd

Figure 5.5: Tag profiles for city core terms associated with different toponyms tags of American cities

5.2 Place tags at the city level

5.2.1 Granularity of place tags

Analysis of tag statistics

As the analysis at the global level showed that the application of terms to describe the urban core does not only vary between culture region, but also between cities from the same language area, the evaluation was refined to the city level. The commonness of different place tags within specific urban environments represented in the bounding boxes of the city-dataset, shown in Table 3.1 was analysed. The original attempt to distinguish between vague/vernacular and official/crisp regions within the city of Zurich had to be abandoned, as the reassessment of colloquial place indications such as 'City' or 'Friesenberg' revealed that there is some form of a (semi-)official counterpart for the vast majority of annotated place names. Also, in the framework of this analysis it was not possible to establish whether users assigned the names of landmarks, such as 'Stauffacher' or 'Hardbrücke', to refer to

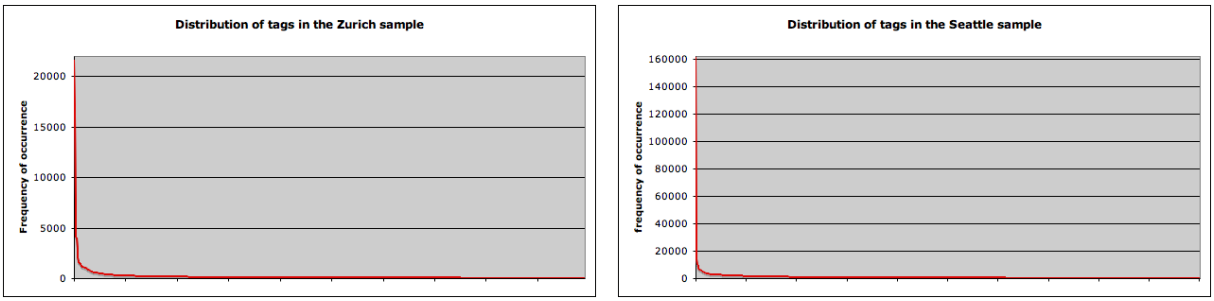
[3]The tag profiles for the Australian data set are shown in Appendix B.1

a wider, vaguely-defined area. As shown in the theory part, the distinction between vague and clearly-defined regions is dependent on the circumstances and the mode of observation (Couclelis, 1996; Montello, 2003) and therefore in practice limitedly applicable. Thus, the evaluation was settled to classify the place tags identified in the frequency lists of the Zurich bounding box into the continental, country, canton, city, district, and the street level. Still, an objective identification and classification of place tags in the entire taglist of the Zurich sample was, despite the established rules, quite challenging. The lower tags were ranked, the more they were susceptible to low quality and idiosyncrasy, which was amplified by the fact that 56.5% of the tags in the list were just used once. All the issues mentioned by Guy and Tonkin (2006) such as extra-long compound tags, misspellings, ambiguity, and singular versus plural forms occurred within the Flickr sample of Zurich. The situation was complicated by the numerous languages applied, sometimes within a single compound tag, reflecting the influence of tourists and the French and Italian speaking parts of Switzerland. The described phenomena yielded a myriad of possibilities to describe a location. Among the 1'000 top-ranked tags analysed for the Anglo-Saxon cities, only a small number of malformed, misspelled, and idiosyncratic keywords appeared. In the Anglo-Saxon tagging world, the language structure is furthermore much less diverse than within Zurich. However, due to referent class ambiguity and ambiguity in the membership of the place granularity level, the numerical values in the tables need to be considered as approximate values of orientation.

tag level	**% of tags**
continent	0.86%
country	13.5%
canton	0.22%
city	18.1%
district	1.22%
street	0.72%
total	**34.6%**

Table 5.3: Proportion of place tags representing different levels of granularity within the bounding box of Zurich

The most common occurring keywords out of 14'046 different tags in the Zurich sample include zurich at rank 1, zürich at rank 3, and zuerich at rank 12. They are all characterised by low coefficients of variation and represent, together with other equivalents of the city toponym (indicated in Table B.1 in Appendix B.2) 18.1% of the tags employed within the bounding box. As shown in Table 5.3, 13.5% of the tags correspond to a place annotation on the country level, while the canton and the continent are less important levels of reference. The place tags designating any of the 'Kreise', neighbourhoods or post code areas sum up to 1.2%. The street level, yielding a portion of 0.72% of tags, might generally be considered as too specific. Altogether, nearly 35% of the tags assigned within Zurich contain some kind of place indication in natural language.



(a) Zurich sample (b) Seattle sample

Figure 5.6: Distribution of 1000 most popular tags and their frequency within the bounding box of Zurich and Seattle

As for Zurich, in all remaining bounding boxes corresponding to the cities of London, Sheffield, Chicago, Seattle, and Sydney the most frequent and ubiquitous of place and non-place tags is by far the official city name itself. The frequency distribution in the samples, as shown in Figure 5.6, confirms the suggestion by Mathes (2004) that tags in online collections follow a power law scenario. If taking the bounding box of a city as a frame of reference, the official toponym is the dominant tag. Regarding the fractions in Table 5.4, in London, the city level is constituted by widely used toponym tags in different languages, reflecting the place's role as a major tourist destination. Chicago is represented by a range of vernacular names with medium ubiquity (both shown in Table B.1 in Appendix B.2). In Sheffield, Seattle, and Sydney no other correct version of the city toponym was identified among the tags in the bounding boxes. The state/country level is generally the second most common spatial reference in users' annotations of photos located in urban space. In Sheffield, for instance, england, yorkshire, and uk appear at the second, the third, and the forth rank. In the Sydney sample, australia, nsw, and newsouthwales are at ranks two, three and four. The substantially higher portions of country/state tags in Sydney and London might be explained by the activity of an abundance of overseas travellers.

tag granularity	London	Sheffield	Chicago	Seattle	Sydney
state/continent	13.13%	6.23%	6.25%	4.65%	17.42%
city	17.72%	12.36%	18.98%	10.88%	14.98%
city core term	0.70%	0.34%	1.28%	0.57%	0.77%

Table 5.4: Proportion of place tags of different granularity associated with georeferenced items within different bounding boxes

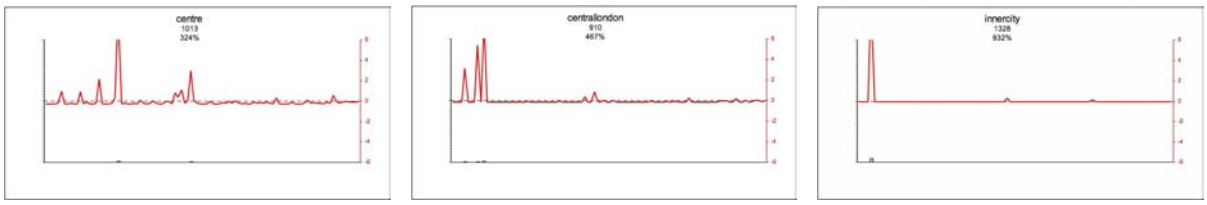
As evident from Table 5.4, the fraction of city core terms is marked by very low values in all cities. As the sample data was mined from the entire extent of the cities, the frequencies of tags at the city level and at the city core level is only limitedly comparable. It is maintainable by the fact that the georeferenced items within the bounding boxes are highly biased towards the centre.

Zurich				**Sydney**			
tag	**number**	**% of tags**	**C. of var.**	**tag**	**number**	**% of tags**	**C. of var.**
city	895	0.353%	195%	city	2'422	0.47%	119%
stadt/altstadt	132	0.052%	487/406%	cbd	1'184	0.23%	256%
(down)town	130	0.051%	550/654%	central	136	0.03%	333%
center/centre	59	0.023%	775/677%	centre	117	0.02%	432%
citycenter	38	0.015%	995%	downtown	112	0.02%	318%
other	12	0.005%					
total		**0.499%**		**total**		**0.77%**	
London				**Sheffield**			
tag	**number**	**% of tags**	**C. of var.**	**tag**	**number**	**% of tags**	**C. of var.**
city	17'698	0.54%	149%	city	261	0.22%	241%
town	1'975	0.06%	318%	centre	106	0.09%	298%
innercity	1'328	0.04%	932%	town	40	0.03%	294%
centre	1'013	0.03%	324%				
centrallondon	910	0.03%	467%				
total		**0.70%**				**0.34%**	
Chicago				**Seattle**			
tag	**number**	**% of tags**	**C. of var.**	**tag**	**number**	**% of tags**	**C. of var.**
city	7'689	0.66%	141%	downtown	6'040	0.40%	121%
downtown	6'039	0.56%	138%	city	2'107	0.14%	167%
town	407	0.03%	531%	center	601	0.04%	194%
center	373	0.03%	176%				
total		**1.28%**				**0.58%**	

Table 5.5: Vague city core descriptions identified among all tags for Zurich and among the 1'000 most frequent tags for the other cities

Even in Zurich, where all central and peripheral neighbourhoods were considered for the classification of tags, the value was not substantially altered (Table 5.3). In Table 5.4, there might be a slight tendency towards a higher proportion of city core tags in larger cities. An exception is London with a comparably low value with respect to its size. In the following paragraph, the terms contributing to the proportion of city core terminology are considered in more detail.

As shown in Table 5.5, in the entire taglist of the Zurich sample, several of the city core terms established in theory were found. However, except for city, generic terms represent only an insignificant number of the tags and were applied by very little different users[4]. 'Zentrum', 'Stadtzentrum', 'Stadtkern', and 'Innenstadt' did not occur, instead, some Anglo-Saxon terminology, such as downtown and centre, was employed to refer to the central area of Zurich. In London, the vague concept of city occurred among the high-ranked tags, but unless applied in conjunction with the City of London, it is not considered to specifically describe the central area of London. As shown in Figure 5.7, the other vague terms occurring among the 1'000 most frequent tags in London were subject to relatively high coefficients of variation. The most ubiquitous was the ambiguous centre. To compare, camden, southbank, docklands, soho, and nottinghill each occurred several thousand times. In the taglist of Sheffield, apart from centre shown in Table 5.5, the unambiguous citycentre occurred 15 times at rank 1150, towncentre only four times. Despite the small sample for Sheffield, the corresponding expressions listed in Table 5.5 have comparably low coefficients of variation.



(a) centre (b) centrallondon (c) innercity

Figure 5.7: Tag profiles for generic place tags within the bounding box of London

In the Chicago taglist, downtown appeared already at rank eleven, while cbd occurred only four times in the entire metadata sample. Instead, all versions of 'The Loop', as which the central business area of Chicago is commonly known, within the 1'000 top-ranked tags summed up to 4'466 occurrences, corresponding to rank 32. Except for the terms displayed for Seattle in Table 5.5, cbd was applied 64 times by a number of different users. Among the top-ranked tags, a striking number of neighbourhood names are present in the Seattle sample, for instance fremont at rank 10, ballard at rank 14, and capitolhill at rank 21. In Sydney, the prevalent generic term is cbd, but also central and downtown are common.

As pointed out above, the counts in Table 5.5 might be biased by the chosen frame of reference, however, that the city toponym is prevalent over lower-level place indications

[4]The tag profiles for the generic city core expressions found for Zurich, London, Chicago, and Sydney are shown in Appendix B.2.

is supported by the sample of Zurich, where district labels from all over the town were considered for classification. The proportion of generic city core terms is generally small and many have high values in terms of the coefficient of variation. More frequent are place names relating to specific neighbourhoods[5]. In larger cities and most parts of the world, 'city' and 'town' cannot be considered to belong to the vague concepts used to refer to the central area of a city, and also 'centre' is ambiguous[6]. If we omit these terms, the fraction of generic city core terms would be further diminished.

Analysis of photo statistics

An analysis of the proportions of photos which have some kind of place tag assigned is considered even more significant than the analysis of fractions of specific tags. For this purpose, the metadata of each instance in the samples of Zurich, London, Chicago, and Sydney were checked if containing any of the above identified place tags from a specific granularity level. This second analysis considers only generic city core terms at the subcity level, also in the Zurich sample.

place tag	London	Chicago	Sydney	Zurich
none	37.0%	32.3%	23.9%	33.1%
country	3.63%	1.75%	8.63%	9.29%
city	38.1%	49.0%	21.2%	19.4%
country&city	19.0%	12.9%	43.2%	35.6%
city core	0.13%	0.11%	0.27%	0.14%
country&city core	0.03%	0.01%	0.03%	0.34%
city&city core	0.83%	2.25%	0.62%	0.50%
all	1.32%	1.68%	2.09%	1.63%
total country	**23.9%**	**16.3%**	**54.0%**	**46.9%**
total city	**59.2%**	**65.8%**	**67.2%**	**57.2%**
total city core	**2.31%**	**4.06%**	**3.01%**	**2.61%**
any place tag	**63.0%**	**67.7%**	**76.1%**	**67.0%**

Table 5.6: Proportion of georeferenced items with place tags representing different levels of granularity

[5]A tagcloud of the 30 most frequent generic and specific place names within London, Chicago, and Sydney is given in Figure B.11 in the Appendix.

[6]A Flickr related-tag analysis (shown in Appendix B.3) revealed that center is mostly used in conjunction with landmarks in New York on the one hand and flowers on the other hand. Centre occurs primarily together with city, downtown, and toponyms from all over the world, indicating that it might quite often be used as a spatial reference within the city. City and town are both employed with downtown but not with centre. City is amongst others associated with traffic and skyscraper, while town relates to church. A preliminary tempting assumption form this analysis is, that users are likely to refer to the urban core by centre, while the remaining ambiguous expressions are applied in many different contexts. Center does not seems to be particularly connected to the inner parts of cities.

From the values in Table 5.6, it is evident that the proportion of photos having place names annotated is even larger than the proportion of tags corresponding to place names within the bounding boxes. Overall, between 63% and 76% of the geotagged photos are associated with a place indication in natural language. This result has to be considered as an underestimation, as specific neighbourhood and district names are not included in the count, also for the city of Zurich. While overall, again the city level is the absolutely dominant frame of reference, in Zurich and Sydney most of the city toponym tags are combined with a tag from the country/state level. As expected from the above evaluation, a very small proportion of images are labelled with a generic city core term only. Users do not often combine a country/state name with a reference to the central district of a city to the same photo, meaning that there is some consistency in the general level of place granularity borne in mind when tagging. Often, people will add the city name together with the generic term, which seems reasonable. This combination is particularly popular in Chicago, were 2.25% of all photos are presumably tagged with chicago and either city and/or downtown. For all other cities, the largest portion of photos associated with generic city core tags has a place indication at all three levels of reference. This is another indication that only users adding a lot of tags, or people with a special focus on place, employ these kind of tags. In total, generic terms to denominate the central part of a city are assigned to 2.31% of the photos in London to 4.06% of the items in Chicago.

5.2.2 Correlation between place tag and georeference

Subsequently, an analysis of the coherence between natural language tags and the geotag accuracy level applied upon referencing images on Flickr is described. It aims to investigate whether the tagging behaviour of users is influenced by the level of detail displayed in the map. For this purpose, the place tag distribution from Table 5.6 was plotted against the geotag level represented in the metadata of the respective photos. The resulting contingency tables in Table 5.7 were analysed by using the χ^2-test for homogeneity. The expected frequencies at the country/world geotag level were sometimes low and are not in line with the rule concerning the restriction of use of the test. Under these circumstances, the null hypotheses of equal distribution could clearly be rejected for all samples, meaning that the granularity of the natural language tags varies between the aggregated geotag resolutions. In order to analyse a possible coherence between the apparently chaotic combinations of natural and formal place tags, the quality and resolution of map detail available in the different parts of the Flickr world map need to be taken into account. The comparison of expected and effective values for the cells of the contingency table of Zurich revealed that the difference is particularly pronounced at the city/region level of accuracy where precise place indications occur more frequently and city toponym tags less frequently than expected. This distribution argues against a particular influence from the map.

For the London and Chicago area, high map resolution is available with detailed district and street name information. For London, the subcity place tags occur less frequently than expected together with the city/region and the country/world geotag accuracy, but are over-represented at the most detailed geotag level. In Chicago, annotations with the

Zurich			
geotag	**street/district**	**city/region**	**country/world**
country	16'003	2'490	60
city	19'869	2'672	63
city core	794	237	0
London			
geotag	**street/district**	**city/region**	**country/world**
country	189'485	24'115	3'702
city	473'288	55'392	9'663
city core	19'073	1'697	249
Chicago			
geotag	**street/district**	**city/region**	**country/world**
country	49'495	2'834	780
city	198'347	13'280	2'288
city core	12'151	906	128
Sydney			
geotag	**street/district**	**city/region**	**country/world**
country	52'552	8'765	15
city	65'950	10'356	14
city core	3'090	329	0

Table 5.7: Relation between natural language tag granularity (country – city – city core) and georeference level (street/district: 16–12, city/region: 11–7, country/world: 6–1)

state name are much less frequent than expected on the city/region level, while the country tags at country/world accuracy are over-represented. As mentioned previously, there was no detailed map information for Australia at the time the data was mined from Flickr, while it has now been enhanced. Only a very generalised view of the coastlines and no zoom level higher than 12 was available for the Sydney area. Geotags applied at a higher resolution had to be made through coordinates or by choosing the satellite image interface. Here, the subcity language tags are clearly less frequent, instead the country tags are over-represented on the city/region geotag level. The findings from Chicago, London, and Sydney imply that there is an interrelation between the geotag accuracy, respectively the level of map detail, and the granularity in represented in the natural language tags.

The relation between formal and semantic georeferences in tags of georeferenced Flickr items leads to the tempting conclusion that the users are partially influenced by the level of detail, the information and the annotation of the map applied upon uploading pictures on Flickr. These findings have to be considered carefully, as the aggregated geotag levels actually comprise a wide variety of map configurations. Furthermore, the analysis is distorted by the proportion of automatically geotagged items represented in the fractions,

which were generated by the use of track-logs, the API or location aware devices, where the user actually does not actually consult the map. Also, tags might be added at a later stage.

5.3 Vague footprints of vernacular place tags

5.3.1 Zurich

As shown in the previous sections, the terms used to describe the central area of Zurich occurred rarely within the bounding box and were, except for city, characterised by very low user ubiquity. However, vague footprints were estimated using KDE, as described in Section 4.2.2, for the four most commonly employed generic place tags within Zurich, that is altstadt, center, city, and stadt. This allowed for a detailed evaluation of the representations and an assessment of the usage of generic place indications in a familiar environment. Furthermore, the approach developed to gain knowledge of ill-defined regions from Flickr tags could be tested, for a city where little information was mined from the database. Despite the very low frequency and the rather high coefficient of variation of 406%, after the removal of x-/y-duplicates in the altstadt sample, 19 points from nearly as many users were left for the density estimation in Figure 5.8(a). Except for three outliers from one user, the points are located where buildings and streets correspond to old town morphology, an area equally distributed on both sides of the river Limmat. Despite the small sample, the area of the old town of Zurich could be delineated by the 50% isoline in a narrower or by the 70% isoline in a wider sense.

The density surface in Figure 5.8(b) represents the 34 instances tagged centre or centre, which are much more concentrated than the referents of 'City' and 'Stadt'. The footprint comprises the part officially called City (shown in the official map of Zurich in Appendix B.5), as well as the area around Zurich Central Station and the old town. The per se ambiguous term is obviously applied by few users in order to refer to the geographically most central part of the city. The point pattern is best approximated by the 50% line, however it cannot be said to reflect a common view and language.

Even though it is an English term, we know from urban theory and personal experience that 'City' is used in German parlance to specifically refer to the central part of a city. This assumption was confirmed by a low coefficient of variation for city within the bounding box of Zurich and a sample that is still based on 308 instances after the removal of duplicate points. The vague footprint in Figure 5.8(c) is focussed on the official neighbourhood named this way, but the general understanding of the City definitely goes beyond the administrative definition. The extent is prolonged along the 'Gewerbeschule' neighbourhood towards the 'Escher-Wyss' area, a recently converted business and entertainment district in the former industrial west of Zurich. It is probably due to its nature considered to be part of the City by some users, even though being distant from the traditional financial and retail district actually named City.

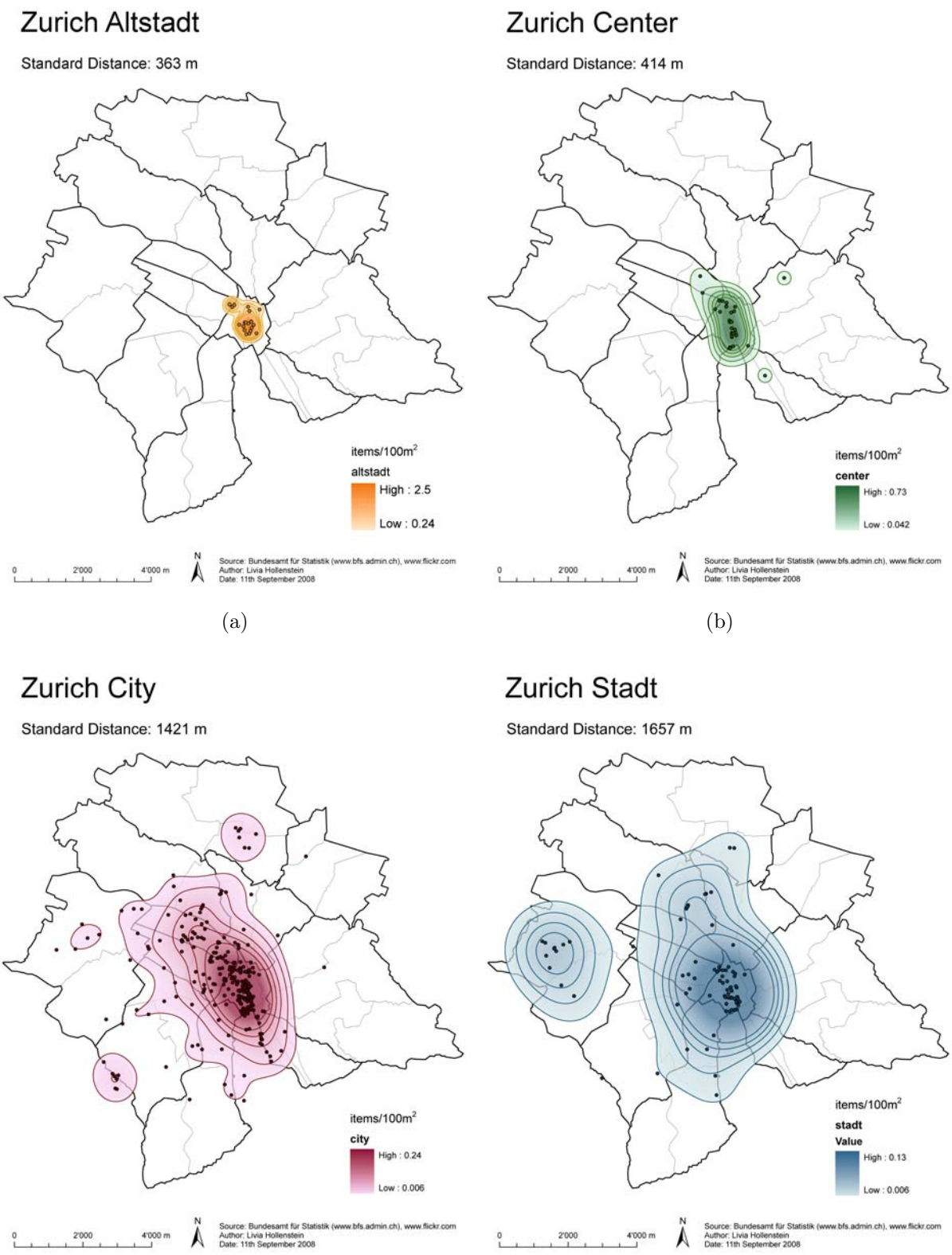


Figure 5.8: Original point pattern, vague footprint and 90, 80, 70, 60, and 50 % isolines for place tags within Zurich

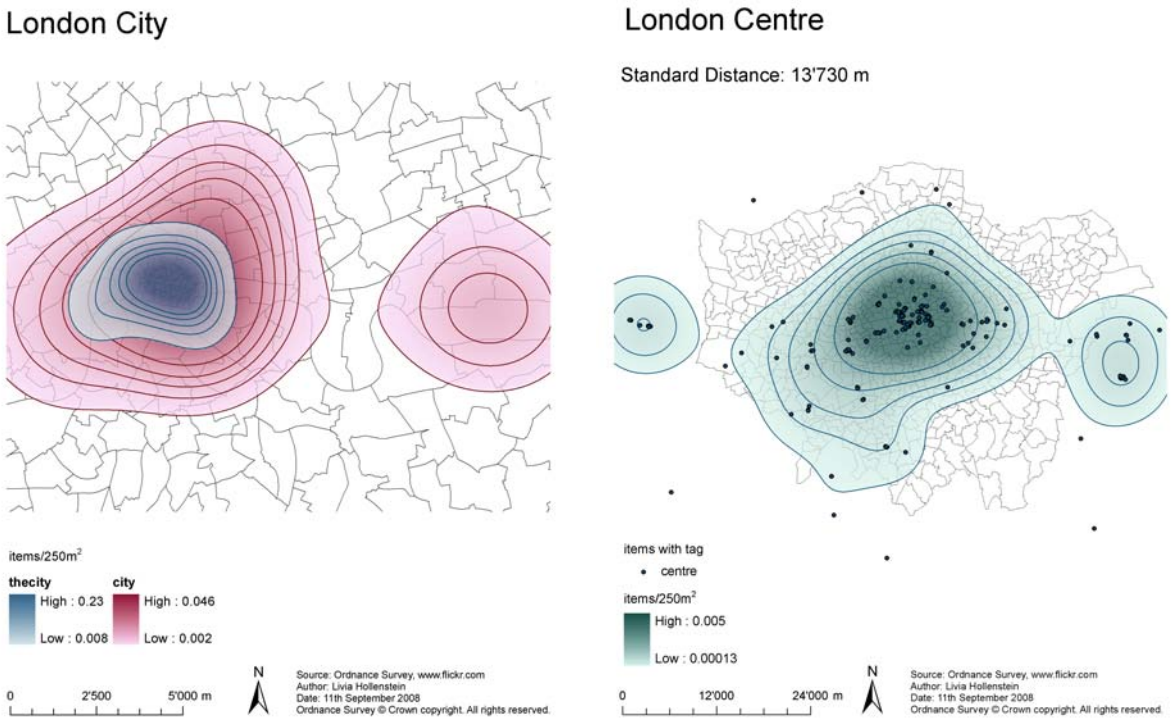
The tag distribution is not extended towards the eastern part of the old town, rather, there is a clear cut in the point pattern at the river Limmat, which is not suitably represented by the KDE in Figure 5.8(c). The outlying peaks are located on the lakeside and the 'Uetliberg', a scenic lookout suitable for photographing the urban landscape of Zurich. Other spurious peak represent 'Oerlikon' and 'Altstetten', two formerly autonomous towns forming sub-centres of Zurich. City is obviously used to refer to the urban core as well as to a wider area with pronounced urban character, clearly distinct from the outskirts of Zurich. The 40% and the 70% isolines are believed to well approximate the narrower and wider understanding of the vague region.

Figure 5.8(d) is based on items annotated with either stadt or town which both have high coefficients of variation. The 88 instances remaining after pre-processing can be expected to have enhanced spatial and user distribution. Still in this case, the technique did not prevent from distortion by a single user who contributed all items responsible for the spurious peak in the western part of the city. Due to a high proportion of outliers in a comparably small sample, not even the 50% isoline is a reasonable approximation of the main point accumulation, which is assumably clustered inside the 20% contour line. Also the standard distance is altered to 1'657 meters which is only slightly lower than the standard distance of 1'972 meters of the total of items within the Zurich sample. As a meaningful basis for comparison is lacking in the sample, it is difficult to determine whether people in general have a wide idea of where the concept of stadt applies within the official borders of the Zurich and additional data is just missing, or whether the conception is commonly bounded to the typically urban inner-city neighbourhoods and this one user is an exception. It can be said that the critical mass is not reached in this sample.

5.3.2 United Kingdom

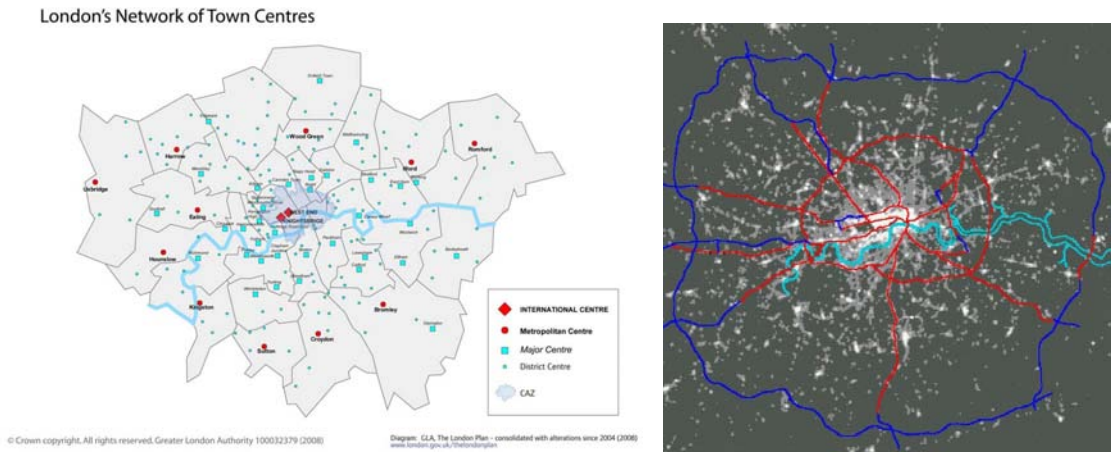
In the bounding box of London there was no expression identified which was popularly used to refer to the central area. The only generic place term with an unproblematic coefficient of variation within the region was found to be city. Inspired by the City of London, City is employed to describe the business core of cities in German language use but it cannot be assumed to specifically refer to the core of major Anglo-Saxon settlements and London in particular. For the special case of London, the difference between the meaning of 'city' and 'City of London', commonly abbreviated as The City, is represented in Figure 5.9(a). The footprints, as estimated from the Flickr tags city and thecity, confirm that users in general accurately distinguish between the spatial indications they use for tagging. The population of city exhibits a standard distance of 5'161 meters which is only slightly more concentrated than the total of items tagged london with a standard distance of 5'759 meters. The city population is not dispersed over the entire area of the city of London, but clustered towards the tourist centre and the Docklands in East London. The 60% contour line of thecity, with a rather low coefficient of variation of 254% before pre-processing, fits the official boundary quite well[7].

[7]http://www.cityoflondon.gov.uk/Corporation/maps/boundary_map.htm, accessed 19th October 2008



(a) (b)

Figure 5.9: Vague footprints for place tags in London



(a) (b)

Figure 5.10: The hierarchical centres of London as identified by (a) the London Plan and (b) Thurstain-Godwin and Unwin (2000)

An inspection of corresponding photos on the Flickr website revealed that the overestimation by trend is mainly introduced by the fact that the high-rise buildings in the financial district are often photographed from a distance. Another reason is the slightly broader conceptualisation of the region assumed to belong to the City, particularly by tourists.

As mentioned in Section 2.2.1, due to its long historic background and its size, many historic town centres constitute a pattern of multiple cores in the London area. The structure is reflected in the London Plan[8], which categorises the hierarchically organised centres of activity within the Greater London Area into two international centres, the West End and Knightsbridge, eleven metropolitan centres, 35 major centres, and 156 district centres, as shown in Figure 5.10(a). This pattern is quite well represented in the approach developed by Thurstain-Godwin and Unwin (2000), shown in Figure 5.10(b), where for all cities, a radius of 300 meters was used to generate density estimations of town centredness from statistical data. The original Flickr point pattern in Figure 5.9(b), stemming from a medium variety of users, looks similar but contains associations to different kinds of centres than centres of urban activity. However, the approach of density estimation pursued here does not particularly well represent the multiple-nuclei structure inherent in the point pattern. As stated by the developers of the HRT extension, the technique chosen for bandwidth determination does obviously not work for multiple-peaked data.

Central London is used to refer to the sections of London which are generally considered closest to the centre. There is no conventional or official definition for the name, but the region has been subject to changing definitions and associations since the 19th century. Colloquially, the region is constituted by the three main sections of the City, the West End, and South Bank[9]. For the purposes of the London Plan, Camden, Kensington and Chelsea, Islington, Lambeth, Southwark, Wandsworth, and Westminster were originally included into the central development area. The planning regions have recently been redefined and a Central Activities Zone was identified, which comprises areas with a very high concentration of metropolitan activities, shown in Figure B.15 in Appendix B.5. The central zone is similar but not identical to more colloquial conceptions of the area which might be established or influenced by the zone of congestion charge or zone 1 of the London Underground (Figure B.15). Due to the critical user ubiquity of centrallondon, the footprint in Figure 5.11, based on about 300 reference points, needs to be considered with care. The region as derived from Flickr comprises parts of all three main sections forming part of Central London, but is clustered in the West End, which can possibly be explained by the district's prevalent popularity. The footprint matches the common definitions quite well, although it is by trend more extended towards the East, particularly compared to the zone 1 of the Underground. The 80% isoline fits the mean of the other definitions best. Users of Flickr seem mostly influenced by the zone of congestion charge in their conceptualisation of Central London.

The footprint of innercity in Appendix B.4, Figure B.12(a), is not representative, as it is predominantly derived from the instances owned by a single user. However, this user,

[8]The London Plan is a planning and development document published by the Greater London Authority on http://www.london.gov.uk/thelondonplan/, accessed 19th October 2008

[9]http://en.wikipedia.org/wiki/Central_London, accessed 21st October 2008

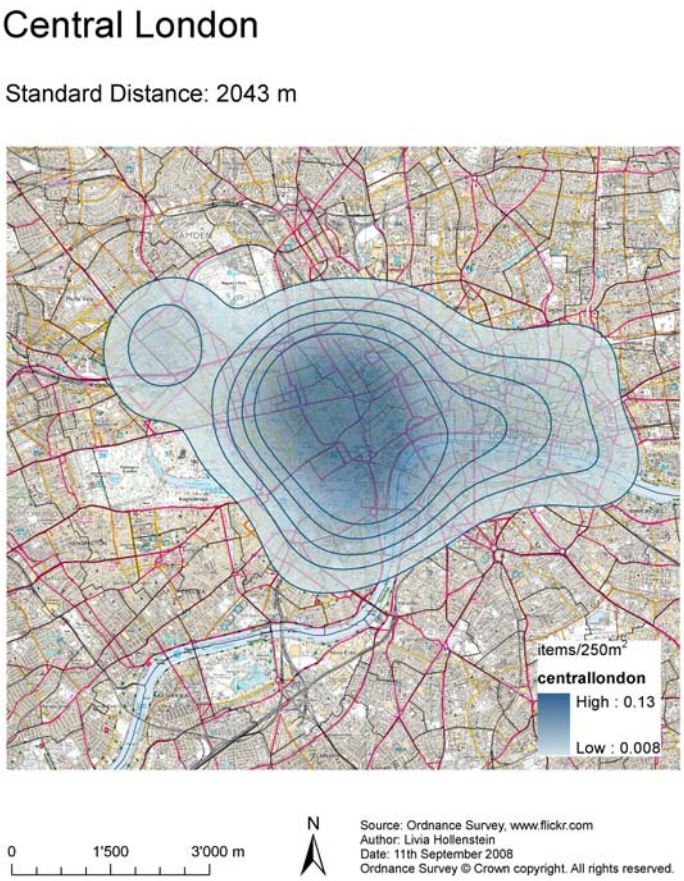
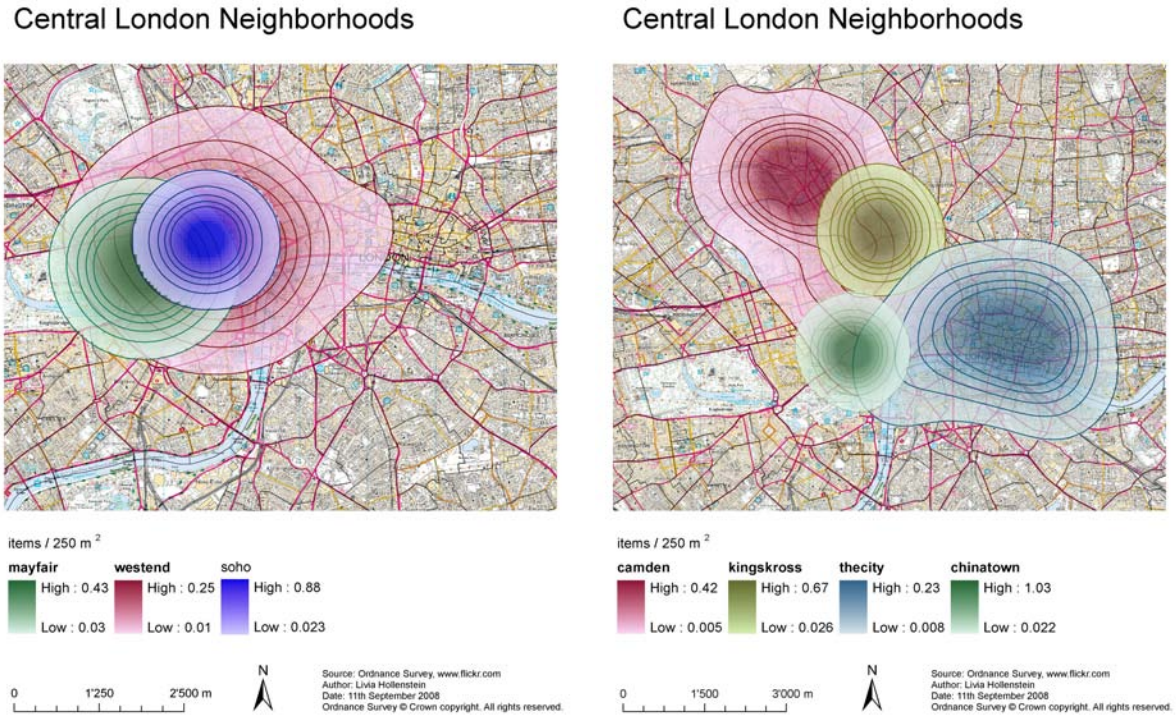


Figure 5.11: Vague footprint for Central London derived from Flickr tags

whose tagging habits were identified to influence the metadata pattern of the tag at the regional and at the global level, has a very distinct view of where the inner-city is located. The portion of the georeferenced photos within London tagged this way are associated with Northeast London in general and the Borough of Hackney in particular, a fact that is also reflected in the global tag cloud for inner-city in Figure 5.3(f). Hackney has a reputation as one of the most multi-cultural but also poor, decayed, and crime-affected regions of London[10]. Clearly, the employment of inner-city in terms of tags includes an assessment of urban space by the user.

As many specific districts and neighbourhoods of London are presented among the high-ranked tags of the London bounding box, the approach of footprint approximation was also tested for specific place names. The places represented in Figure 5.12 were chosen by means of a London travel website[11] to ensure vernacular usage. The footprints are based on between 212 instances for mayfair to 2'230 references for camden. They are considered reliable as exhibiting low coefficients of variation, with a maximal value of 271%

[10]http://en.wikipedia.org/wiki/London_Borough_of_Hackney, accessed 21st October 2008
[11]http://golondon.about.com/od/planningyourtrip/a/geography.htm, accessed 9th August 2008



(a) (b)

Figure 5.12: Vague footprints of vernacular areas of Central London

for westend. Detailed discussion of the results in Figure 5.12 is omitted while generally, it can be said that the arrangement and conception of the footprints meet the conventional and official definitions with astonishing accurateness. Several point distributions associated with vernacular place names have their centroids exactly on the respective annotations of the Ordnance Survey map, while the text on the backdrop used on Flickr is slightly displaced. The majority of places are accurately represented, such as the district of Mayfair, which is conventionally known to be roughly bordered by Oxford Street to the north, Regent Street to the east, Piccadilly and Green Park to the south, and Hyde Park to the west[12]. This extent is covered by the 50% isoline of the mayfair footprint. Generally, the 50% volume contour delivers the best approximations of the highly correlated point patterns of the smaller regions, while for more extended areas such as the West End the 70% or 80% isoline, or even the 90% contour for Camden are believed to be more appropriate. Minor problems occurred in the metadata pattern of Camden[13], which is internally distorted towards the popular Camden Town and the St. Pancras area, as well as in the pattern of Chinatown, which is overestimated due to outliers. Generally, the areas are better represented by the underlying point patterns than by the density surfaces, which are round overestimations of the assumed extents. In Section 2.2, the point was made that people are likely to identify the boundaries of urban regions by means of major

[12]http://en.wikipedia.org/wiki/Mayfair accessed 15th Oct 2008
[13]http://www.camden.gov.uk/ccm/content/global/onecolumn/camdenmap.en, accessed 15th Oct 2008

roads, train tracks or park boundaries. This is confirmed for some of the point patterns produced by the users but not reflected in the density surface representations.

Approximations of the footprints of South, East, West, and North London are mapped in Figure B.12(b) in the Appendix. The representations are based on 328 to 2'389 reference points and coefficients of variations of 238% for westlondon to 427% for northlondon. These kind of place indications are apparently less common than specific place names, while enjoying higher popularity than generic city core terms in London. The overall configuration of the areas makes sense, while they are highly clustered and overlapping at the geographical centre of London. Also the footprints for city and citycentre in the bounding box of Sheffield are shown in Appendix B.4. The estimations of the density surfaces are based on average commonness of tag usage but worked quite well despite the small samples (52 for citycentre and 129 for city). The 50% contour of the citycentre estimation closely matches the boundaries derived by human subject tests, but are again more circular compared to the elongated shape established by Mansbridge (2005). An inspection on the Flickr website revealed that the ambiguous centre is mostly used to refer to the city centre within Sheffield. The standard distance and the footprint of city, although distorted by outliers, is only slightly more scattered than the metadata pattern of citycentre. Also in Sheffield, city seems to be used for a wider area than the actual centre but rather for more pronounced urban districts than for the outskirts.

5.3.3 North America

The CBD of Chicago appeared repeatedly in urban theory and was referred to as 'The Loop' already by Burgess in the 1920s. This place name is also prominent on Flickr, where the total of the tag loop occurred 1'340 times within the 1'000 top-ranked tags of Chicago, yielding an original coefficient of variation of only 171%. Cbd on the other hand, occurred only four times within the data set. The Loop was, for instance, defined by Johnson in 1941 as being bound by Roosevelt Road to the south, by the Chicago river to the north, by the south branch of the river to the west, and by Michigan Avenue/Beaubien Court to the east Murphy (1972). This definition is adopted in the backdrop maps of Zillow, while the official delineation, shown in Figure B.5 in the Appendix is less far extended to the south. The footprint derived from Flickr in Figure 5.13(a) roughly matches the delineation of the Zillow neighbourhood. Though, it is slightly deflected to the north over the Chicago river, even though the latter is supposed to mark a clear boundary in the cityscape. This is partly induced as many users take photos of the skyline of the Loop from the John Hancock Center, which is located north of the river. The neighbourhood of The Loop is best approximated by the 80% contour line, but again, the shape turns out round in relation to an environment marked by an orthogonal street pattern. The footprint of the downtown in Figure 5.13(b), for which no official counterpart exists, is located in the same place as The Loop but extends much further which would be in agreement with the theoretical considerations on the more comprehensive function of the American downtown.

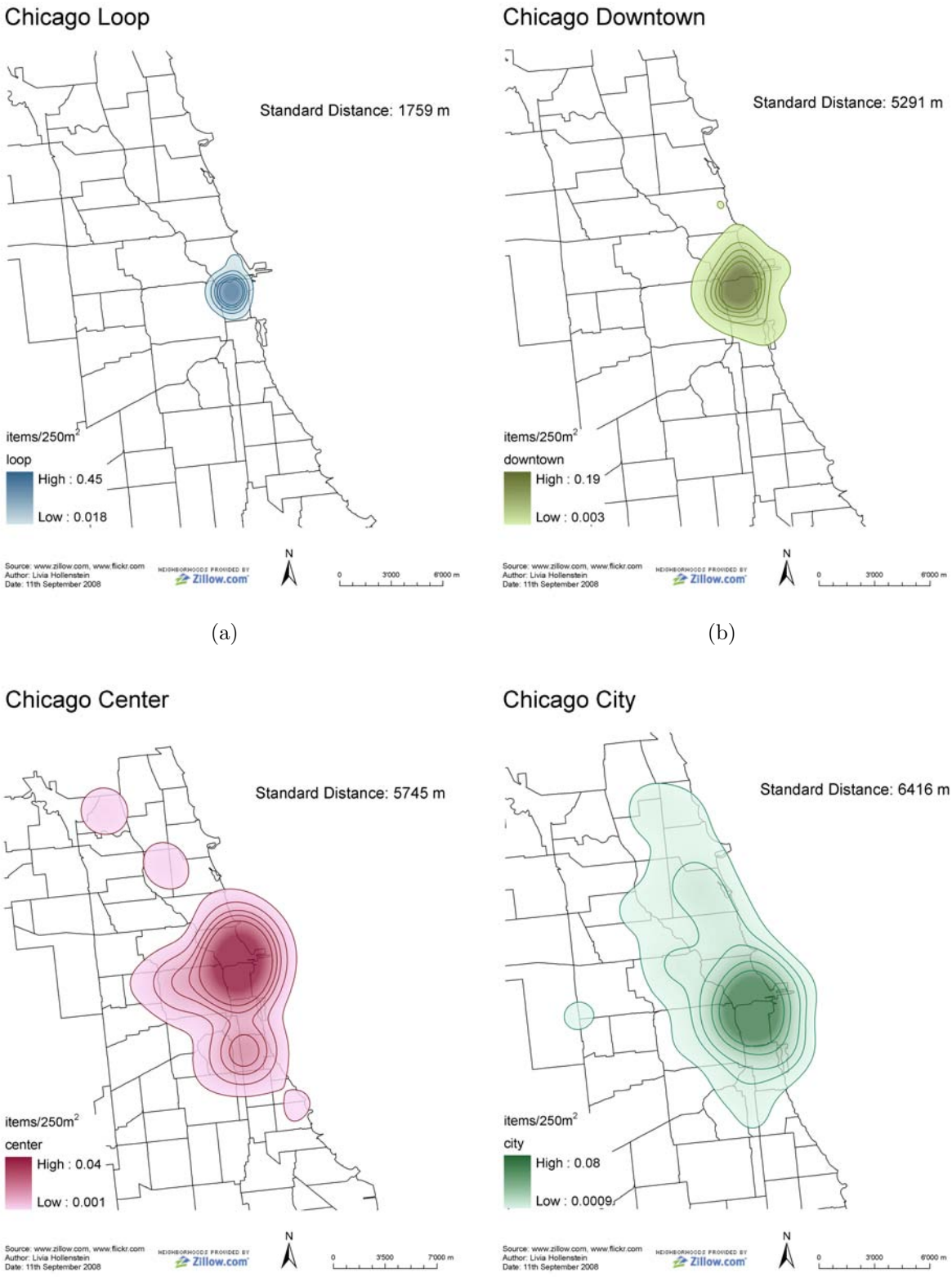


Figure 5.13: Vague footprints for tags in Chicago

The metadata pattern of center, represented in Figure 5.13(c) is more extensively scattered and based on a smaller point sample than downtown. It might be used to refer to the city centre, but occurs frequently in conjunction with buildings and institutions, such as the IIT Campus Center in the south, the United Center in the west, and the Park Village Nature Centre in the north of Chicago. A random inspection revealed that even the photos geotagged at the geographical centre are related to buildings and landmarks located downtown, such as the John Hancock Center, the Chicago Cultural Center, and the Harold Washington Library Center. The problem arises due to the inconsistent term boundaries of user-generated tags. Only a small fraction of users are actually referring to the city centre, all of whom turned out to be non-American residents. The point references for city in Figure 5.13(d) are insignificantly more clustered than the items tagged chicago. As the best approximation of the point pattern is given by the 90% boundary, the term is not considered as being used to refer to the most central part of Chicago.

Seattle has a reputation for being 'a city of neighbourhoods', although no official definitions of neighbourhood names and boundaries exist and the designations of districts at the sub-city level are disputed[14]. Due to the different ideas about the configuration of the neighbourhoods and due to the constantly changing meaning of district names, the Seattle City Clerk's Office has designed the 'Seattle Neighborhood Map Atlas'[15]. The atlas, based on a variety of semi-official documents and planning studies, was not intended as an 'official' map, but was developed in order to improve the indexing and retrieval of documents with place names in the City Clerk's Office and the Seattle Municipal Archives. The Seattle downtown area as delineated and named by the City Clerk's Office is shown in Figure 5.14. Downtown was found to occur 6'040 times with an original coefficient of variation of 121% in the bounding box of Seattle. The footprint generated from this broad perception in Figure 5.15(a) is larger than the area denominated by the semi-official map and the boundaries marked by Zillow. The point pattern, which is widely scattered, is best represented by the 70% or 80% isoline. Photos of the Seattle skyline taken from the top of a hill and from the other side of the bay led to spurious peaks in the footprint. To enable a more in-depth evaluation of the central area, a footprint of the CBD and of the neighbourhood of Belltown was derived from the Flickr point data. Belltown had a coefficient of variation of 238% before pre-processing and the footprint matches the definitions of the reference maps quite closely. As the sample for the CBD in Figure 5.15(b) is based on only 34 points, the footprint is malformed due to outliers. If a threshold was set at the 50% contour, which represents the majority of points best, we would end up with a configuration of the central area derived from Flickr tags that converges close to the map by the City Clerk's Office. All the instances tagged center within Seattle were located around the popular entertainment centre featuring Space Needle, one of the main landmarks of Seattle and are therefore not considered as referring to the city centre.

[14] http://en.wikipedia.org/wiki/Seattle_neighborhoods/, accessed 18th Oct 2008
[15] http://clerk.ci.seattle.wa.us/public/nmaps/, accessed 18th Oct 2008

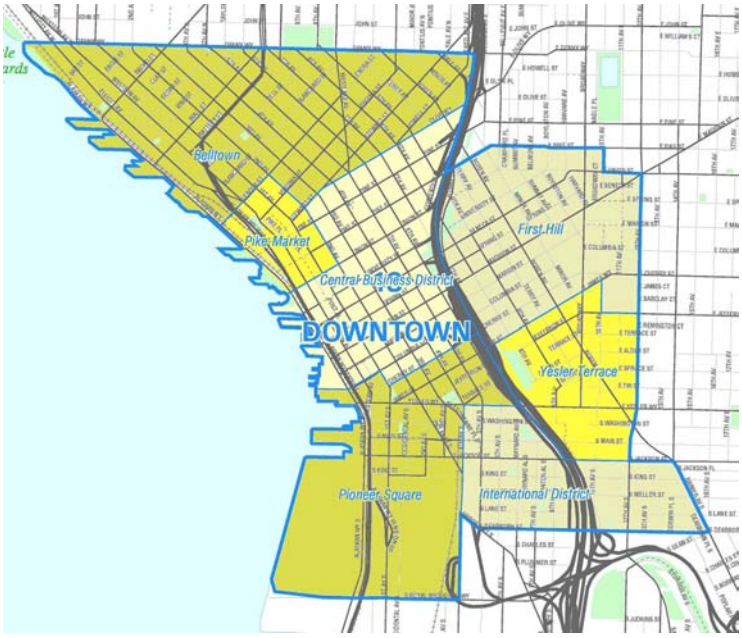
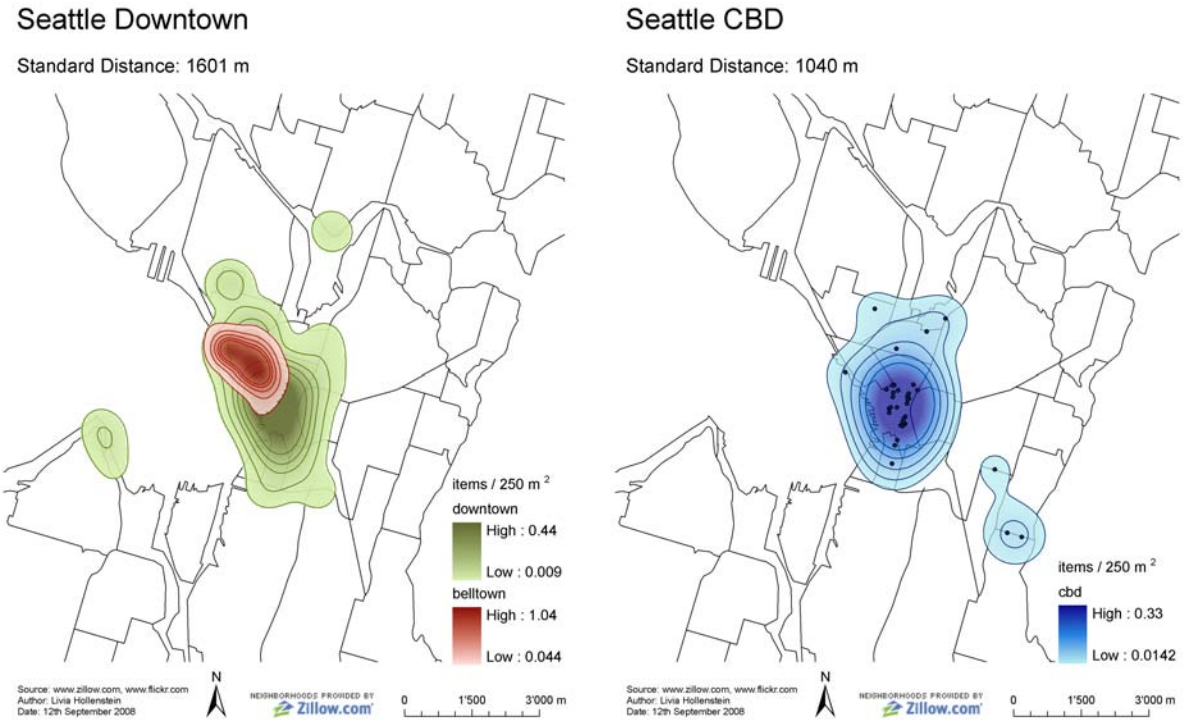


Figure 5.14: Semi-official districts of downtown Seattle defined by the Seattle City Clerk's Office (Source: http://clerk.ci.seattle.wa.us/public/nmaps/html/NN-1240L.htm)

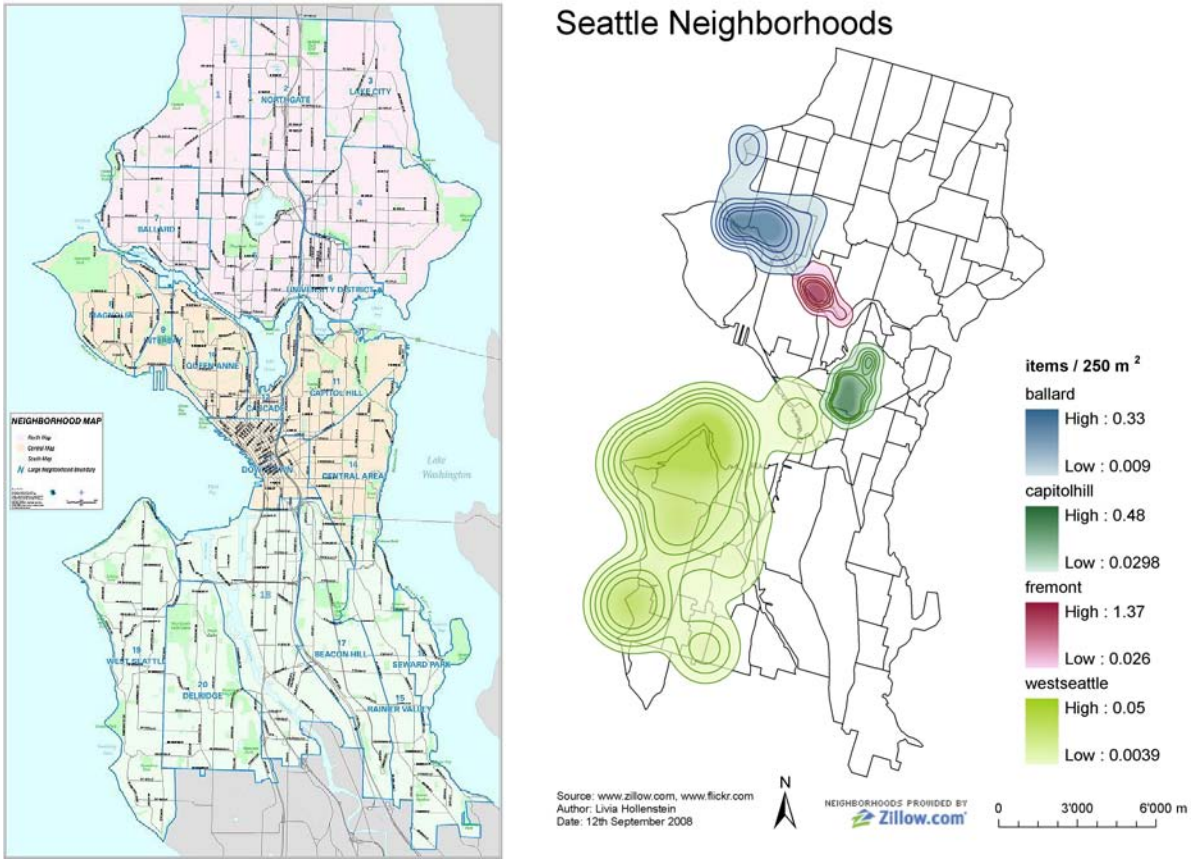


(a) Downtown neighbourhoods from Flickr

(b) Central Business District from Flickr

Figure 5.15: Seattle downtown area as derived from Flickr tags

The importance of neighbourhoods in the city of Seattle was reflected in the high-frequency tags of the taglist. Specific names of neighbourhoods were frequently and commonly employed in terms of tags within the bounding box. The minimal user ubiquity of the neighbourhoods represented in Figure 5.16 was established for westseattle, with a coefficient of variation of 202%. Generally, the footprints as derived from Flickr show broad agreement with the boundaries in the reference maps. The footprint for Capitol Hill is clustered towards the actual hill and is more congruent with the narrower definition of the area in the Zillow map than with the area denominated by the Clerk's office. The Flickr representation of Ballard on the other hand is closer to the delineation of the Clerk's office. Generally, the users seem to have a wider impression of the area belonging to West Seattle than established by the reference maps, if taking the closest approximation of the point pattern, which is given at the 70% contour line. Overall, the performance of footprint approximation from georeferenced tags varied and will be discussed further in the following chapter.



(a) (b)

Figure 5.16: Seattle neighbourhoods according to the City Clerk's office (a) (Source: http://clerk.ci.seattle.wa.us/∼public/nmaps/fullcity.htm) and as modelled from Flickr tags (b)

Chapter 6

Discussion

This thesis project was motivated by problems of information retrieval and the emergence of the need to incorporate common-sense geographical reasoning and knowledge into the design of systems and spatial representations. The literature review has disclosed the wider implications of the problem in the field of geography, including questions of spatial cognition, location awareness, behavioural geography, and urban development which provide the background of the research questions forming the basis of this study. In the two concluding chapters, the previously introduced results are discussed with respect to the research questions formulated in Section 1.2, and implications for previous and future work are considered.

6.1 Description of geographic space in user-employed tags

The need to consider people's colloquial reasoning about space in order to enhance Geographic Information Systems and Services, used by a growing community in professional and daily life, is widely acknowledged (Egenhofer and Mark, 1995; Couclelis, 1998; Montello et al., 2003). It has frequently been stated (Arampatzis et al., 2006; Jones et al., 2008; Twaroch et al., 2008) that people are likely to employ vague terminology such as 'downtown' or 'Midwest' in daily communication, as well as when using information services and search engines. In order to investigate people's intuitive reasoning about urban space, this study has drawn on the abundance of absolute references between places and associated descriptions available from georeferenced labels in Flickr, which is considered as a case study for tagging systems in general. Against the background of a globally comparable and multilingual database, the research problem has been formulated as follows.

How do people describe urban places in terms of tags depending on language and culture region?

Within the scope of this thesis, the problem has been addressed by the exploration of

generic place descriptions used in the Anglo-Saxon and the German speaking world to refer to cities' geographic cores, the prototype of a vague geographic region. The data-driven approach has been guided by theory of urban geography, providing the background as to the origin of the generic terms describing the urban cores.

6.1.1 Usage and meaning of generic city core terms

By investigating the properties of place tag distribution as well as the co-occurrence and frequency of keywords within different cities, it could be shown that the generic city core terms identified in the literature are colloquially used. However, they are applied to a very different extent and with variable meaning and popularity in diverse parts of the world. Patterns of spatial distribution for different generic place tags in the world maps were generally confirmed but also refined by the analysis of users' provenance, i.e. assumed native language, and by the tag clouds representing co-occurrence. The latter turned out to provide easily extractable and similar information as depicted by the world maps. In this form, tag clouds should ideally be applied to give spatial meaning to non-geocoded data. The technique could have been adopted to analyse the spatial relatedness of non-georeferenced items by evaluating the most frequent toponyms within such data sets by means of tag clouds.

As assumed in literature, downtown is the most popular term in all of the cities representing the US, exhibiting a bias in user ubiquity of only 52%. Even though the central neighbourhoods of non-American cities are supposed to exhibit a substantially diverse appearance, downtown was identified as the prevalent Anglo-Saxon city core term on pretty much every continent. Obviously, the expression has expanded over the world or rather, it has been dispersed together with people in view of increasing mobility, tourism, and transnational migration. Due to the high variation in the employment of the tag among users who only posted few photos within the UK, it might be assumed that downtown is mostly used by tourists to describe the centre of British cities. On the other hand, at a global level, more users indicating a British location of residence used downtown than city centre as a tag. CBD is commonly used in Australia and New Zealand, and occurs in major Asian cities, but is definitely not (anymore?) part of most people's vocabulary in North America, as stated by Murphy (1972) and Fogelson (2001). Neither is it present in the UK at all. From the two examples visualised through footprints, it could be concluded that the place corresponding to the concept of the CBD, regardless of its name, is considered as being part of, but not covering all of the downtown. This is in agreement with the theoretical considerations about the extended function of the American downtown compared to the mono-functional CBD (Fogelson, 2001).

As expected, city centre is mostly used in the UK. Though, it seems only limitedly appropriate for the multi-centred urban structure of the Greater London Area. The term is applied to the central parts of smaller settlements such as Sheffield, where it is often being abbreviated to 'centre'. The analysis of tag clouds, users' provenance, and photos tagged in Chicago revealed that city center is not often used among Americans but rather among non-native speakers. The activity of tourists might therefore also explain the tag's

low bias in user ubiquity (116%) in the American culture region. The notion of 'central' is used in conjunction with some of the major cities in the world, such as New York, Beijing, and London. However, within London the expression yields a surprisingly low popularity as regards user ubiquity. Besides, it can be considered as a rather academic concept. Also inner-city is seemingly an almost purely scientific expression. The georeferenced instances tagged with inner-city at a global level stem basically from one user living in the US and one user living in London. It is mostly used to refer to the rather poor and deprived neighbourhoods of northeast London which suggests an explicit connotation and assessment of urban space in the employment of the term.

As implied by its size and importance, London is considered as a special case with limited significance for the British linguistic region. Further comparison for the German culture region was not established and the usage of terms with respect to their spatial applicability within Australian cities was omitted in view of the scale of this thesis. Except for a preliminary statement about the relation between the American downtown and the CBD, which could be considered as a subregion of the first, the somewhat unlucky choice of reference cities in this context did not allow to draw conclusions about the universal dimension and meaning of the terms. Due to the strong place-dependency observed in the employment of tags, the observations about the nature of colloquial place indications in user-generated annotation, as discussed in the following section, are considered more significant and relevant than the universal use of generic place tags.

6.1.2 Characteristics of user-employed place indications

As mentioned earlier, the distinction between vague/vernacular and official/well-defined had to be omitted due to the lack of a consistent theoretical definition and problems in the layout of the empirical investigation. If being confronted with a list of place tags, the question about the nature and degree of vagueness and 'vernacularity' cannot be answered concludingly. While vagueness is formally defined, Montello (2003) discriminates between administrative and cognitive regions, but shows how the two categories tend to be intermingled at the same time. While administrative regions are the only type of regions with the potential for truly crisp boundaries, they are only well-defined if used within these legal terms, for instance to collect and evaluate statistical data. People referring to the place name of an administrative region are not likely to mean the exact extent of the legal definition, as there are always parts which are considered as more typical representatives of a specific place (Fisher, 1996; Campari, 1996; Montello, 2003). As mentioned by Hill (2006), authorities in charge of place naming often adapt and formalise colloquially used place names. That the existence of official definitions does not prevent a region from being vague was also shown by Byrkit (1992: 6), who stated that "... the United States government has almost as many "Southwest" definitions as there are agencies and departments within the bureaucratic colossus". The observations made by other authors and within the completion of this project indicate, that vernacular/vague and legally defined are not exclusive, meaning that many regions are subject to contested rather than to conceptual vagueness and that it is not always possible to differentiate between the objective and

the subjective component of uncertainty. Hence, the nature of a region is determined by political, social, and cognitive processes on the one hand and the mode of observation and representation on the other hand (Couclelis, 1996). However, within the scope of this thesis, the classification of place tags into different levels of granularity and generality, that is into specific and generic place names, revealed interesting insights into the nature of intuitive and common-sense reasoning about urban space.

The analysis of georeferenced taglists associated with generic city core tags and of the users who contributed to the data sets revealed that tags such as innercity and citycenter are employed by only a small group of users. Innercity, central, cbd stem from taglists with a mean of 15.2 keywords per picture. Even the popular downtown is associated with photos having on average 11.8 tags. This is way above the average of 5.2 and 7.0 tags, respectively, as calculated for the reference samples in this project and the mean of 7.1 keywords computed for a georeferenced Flickr data set by Wood et al. (pers. comm.). Thus, these kind of terms are probably assigned by people describing elaborately and in great detail, who probably have a special awareness and interest in urban space and/or a special focus on photography of urban landscapes. Also when taking single cities as a framework for identification, classification, and quantification of tags, only a marginal fraction of the high-ranked keywords correspond to generic place tags. Except for downtown in the US and CBD in Australia, these kind of tags tend to originate from a small range of users. Despite the different forms of analysis employed in the study, the application of the concept of 'city' as well as 'town' remains uncertain, in particular with regard to the Anglo-Saxon world. In most cases it could not be established if they are used to refer to a specific part of a city or if they are rather assumed to be general concepts of the environment. If excluding these two terms from the counts contributing to generic place tags, the fractions would be reduced substantially, on average about a factor of ten. In order to get an actual estimation of the perceptual centredness within a city, the restriction used for filtering the Flickr data should have been chosen differently. For instance, frequently occurring place tags in a presumably central section of a city could have been combined to calculate aggregated density surfaces.

Apparently, the larger the city, the higher is the proportion of tags used to designate its central part. An exception is London, for which, despite the size or perhaps just because of the size, there is no widely acknowledged consensus on a means to refer to its central area. Generally, it could be shown that the employment of terminology strongly depends on the city under consideration. CBD within the US, for instance, is mostly used for the business district of New Orleans and to some extent for Seattle, both cities for which there is a specific (semi-)official definition corresponding to the place concept. The strong interrelation between official naming and vernacular terminology has also been pointed out by Hill (2006). The generally marginal occurrence of generic place indications is in agreement with Sanderson and Kohler (2004) who found that 'north', 'south', 'east', and 'west' were rarely used in a vague directional sense within web queries, but rather as part of place names or institutions. In this sense, also 'central' might be thought of as part of a place name within specific cities, rather than a spatial concept of centrality and superior accessibility.

Specific place names belong to the most frequent keywords used within the cities. Regardless of cultural and linguistic backgrounds, they occupy many of the top-ranked tags among the several hundred thousand distinct keywords occurring in the bounding boxes. The prevalence of keywords corresponding to places is particularly striking if considering tag statistics per photo. An overall mean of 68% of georeferenced photos have a place indication in natural language associated, even if neglecting specific neighbourhood names. Without any exception, the official toponym of the city (or its English equivalent for Zurich) is the absolutely prevalent keyword in the populations of georeferenced tags. The city toponyms are followed by place tags referring to the superordinate level such as countries and states. The comparison of exact values might be biased by the chosen frame of reference and it is an obvious drawback of the analysis that the count at the sub-city level is restricted to generic city core terms. However, that the city toponym is prevalent over lower-level place indications is supported by the example of Zurich, where all district labels were considered. These findings are in agreement with Rattenbury et al. (2007), who identified sanfrancisco as the dominant tag in the San Francisco Bay Area, and Wood et al. (pers. comm.), who found london on the first rank, while england, respectively the uk, attained rank 6 and rank 11 with regard to user ubiquity for georeferenced items posted within the UK. Obviously, the city level is also essential when seeking information. Zhang et al. (2006), who analysed about 400'000 web queries containing a place name, found that about 84% of the place indications belonged to the city level while only 16% referred to a state/country. The basic geographic level people intuitively think of when describing the location of online items is undoubtedly the city name. Perhaps they consider the lower scale regions within a city as too specific to be searched for by others.

Place names relating to specific neighbourhoods are not as prominent as tags at the city and country level but occur quite frequent in larger cities. In the bounding box of Sydney, for instance, newtown attains rank 54, therocks rank 60, and glebe rank 64. The prominence of neighbourhood names is even more pronounced in Seattle, where fremont is at rank 10, ballard at rank 14, and capitolhill at rank 21. For London, the 53 vernacular neighbourhoods listed on a travel website[1] occur 26'623 times altogether, yielding a portion of 0.8% of the tags in the bounding box. A striking example is the central business district of Chicago, which by theory and function definitely corresponds to a CBD, but is referred to as such only four times. Users of online tagging systems do not label the place by its function, but by its specific name, 'The Loop'. The prevalence of specific place names over generic place concepts and the strong place dependency of terms, which are implicitly generic, imply that people rather think in terms of places than functional entities when referring to location in common language. In the intuitive process of user-contributed tagging, place references in form of abstract, functional concepts are little thought about or considered limitedly meaningful. The users' tagging behaviour rather reflects what people have learned about specific, named places in their environment by interacting with other residents, official authorities, and by using maps (Montello, 2001).

Initial results of a simplistic analysis concerning the accuracy level employed upon posting a photo on the map and the semantic granularity applied in describing the location suggest

[1] http://golondon.about.com/od/planningyourtrip/a/geography.htm, accessed 9th August 2008

that there is some interrelation between the two tagging modes. However, the assumed correlation needs further investigation as it might be distorted by the photos which were geotagged automatically by the use of track-logs, the API, or location aware devices. Due to the layout of the analysis, the reasons for the suggested interrelation remain obscure. Apart from the possible influence by the nature of the map used for posting, also the affiliation of a particular group of users might influence the distribution. It can be imagined that owing to special interests in terms of photography, a general attitude due to the cultural background, pronounced orientation ability or familiarity with a city, some users are likely to put a special effort in the process of georeferencing as well as in the assignment of semantic place tags, and therefore cause the coherence in the two tag modes. A preliminary analysis by Girardin et al. (2008) in the city of Rome suggested, for instance, that users from Spain tend to provide less accurate spatial metadata than a comparison group from Germany.

6.2 Tagging systems for capturing vernacular geography

The potentials and drawbacks of user-contributed content in online tagging systems in terms of information organisation and retrieval (Golder and Huberman, 2005; Winget, 2006), the generation of ontologies (Schmitz, 2006), and its suitability to represent the perception of the individual (Guy and Tonkin, 2006) as well as distributed knowledge (Weiss, 2005; Steels, 2006) have been discussed at length. As regards the spatial component of user-contributed metadata, Rattenbury et al. (2007) showed that tags representing places exhibit meaningful spatial correlation. Grothe and Schaab (2008) have successfully modelled large-scale vague regions from georeferenced Flickr tags. The approach of gaining knowledge about vernacular regions from the Flickr database is straightforward as, once the data is collected, it provides formal coordinates to a place and a direct link to users' associations therewith. In this project, different aspects of the global data reflecting local knowledge have been investigated in order to answer the second research question.

Is user-generated metadata in online photo-collections suitable to capture vernacular geography?

Vernacular regions are said to be cognitive regions, which are shared among many individuals of a society (Montello, 2003). This thesis project has focussed on the nature of place tags and the conditions required to derive a collective view from the abundance of Flickr data. Also, the reasons leading to errors and distortion in user-contributed data were investigated. As pointed out in answer to the first research question, georeferenced metadata from Flickr is a rich source of information about geographic regions, as specific place names are popularly used to attribute the characteristics of georeferenced photos. The evaluation of the results and underlying data revealed that, in order to capture a distributed spatial cognition, some important observations need to be taken into account if relying on information from georeferenced tags in online photo-collections.

All the shortcomings characteristic of user-contributed metadata mentioned by Golder and Huberman (2005) and Guy and Tonkin (2006) were found in the samples mined from Flickr and made the manual identification of tags a somewhat difficult undertaking. The main problem regarding the chaotic nature of tags was inconsistency with respect to term boundaries and subsequently ambiguous meanings of keywords. Not even for a human classifier it is possible to infer the intentions behind a tag such as `center`, for which checking of the context revealed that it was mostly part of the name of some convenience centre and did not refer to the centre of urban activity. Ambiguous keywords could probably be disambiguated automatically by considering co-occurring tags, but the approach presented here is most straightforward if working with unambiguous or specific place tags. Even though about 50% of tags were identified as unique keywords, the likeliness of idiosyncratic labels is substantially lowered with increasing level of occurrence and particularly with increasing levels of user ubiquity as modelled by variance.

If considering place tags with a proven common social meaning and significance, locational error is mostly induced by the formal georeferences. Some of the users do not seem willing or able to correctly locate on the map. As expected, a higher accuracy level results in slightly better quality of geotags, a fact which can be taken advantage of in the process of footprint modelling. A more essential problem is the spatial bias in the Flickr data. Georeferenced photos do not equally cover the whole extent of a city, but are, particularly in smaller settlements such as Zurich and Sheffield, clustered towards the geographic and cultural centre of activity. Furthermore, picture locations in all cities are distorted towards 'photogenic' subjects, sightseeing attractions, special landmarks, and the waterfront. This shortcoming in the Flickr data was also observed by Grothe and Schaab (2008) and referred to as 'first order effects'. As Jones et al. (2008) state, also the results of web-mining techniques are influenced by the occurrence of place names on the Internet which is biased towards places with higher population or popularity. The incomplete coverage of the web as a whole can be reduced by a careful choice of search queries (Jones et al., 2008). On Flickr, the problem seems more pronounced, as it was not eliminated even by removing all point multiples.

The Flickr database contains an immense amount of spatially relevant, empirical data and keeps growing at a tremendous pace. However, the availability of a huge sample compared to manually collected information from human subject tests does not yet ensure the representation of a collective perception. Not even a sample of 1'061'883 photos with 1'328 occurrences of a distinct tag prevents the bias through a single user, as it was the case for inner-city in London. The fact that given the technical possibilities, a few prolific users are able to produce a lot of data and significantly distort the information, needs very careful consideration. The generation of tag profiles, as suggested by Wood et al. (pers. comm.), has proven a valuable means of accounting for and possibly disregard the contributions of single users. From the experiences made within this project, an overall coefficient of variation of about 300% could serve as a benchmark for distributed cognition and collective knowledge.

Luckily, the rich set of metadata associated with every photo allows the handling of bulk uploads and erroneous data. When factoring the above mentioned constraints into the

generation and evaluation of results, the data has been shown to match the areas of public parks in London quite accurately. Items were also highly correlated with regard to less clearly defined places at the sub-city level. The footprints related to neighbourhoods and districts are generally based on the impressions of a lot of different users. The metadata patterns checked for London and Seattle fit the official and conventional definitions of neighbourhoods surprisingly well. This is a clear indication that the average user has distinct knowledge and ideas about places, their names, and their extent. For certain point samples representing neighbourhoods in Zurich and London, it is even evident how discontinuities in the urban landscape such as major roads or park boundaries are used as delimiters of an area. Furthermore, the mapping revealed that most of the users take care in tagging. In this sense, Winget (2006: 13), stating that the majority of users "at the very least have the best intentions" when tagging their images, can be confirmed.

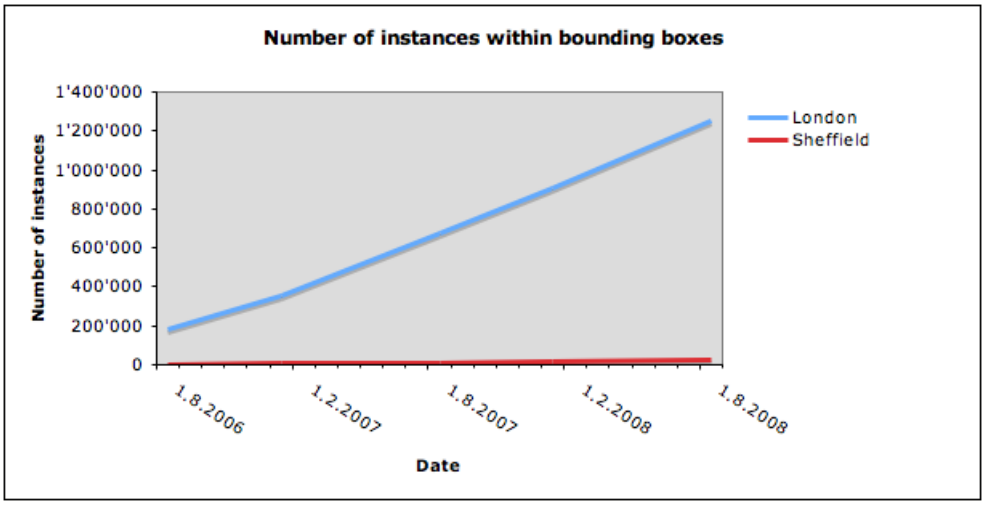


Figure 6.1: Evolution of the number of geotagged instances in the bounding boxes of London and Sheffield

Not only the data, but also the technical environment of Flickr is subject to constant adjustment and change. As it became obvious from the issue of the backdrop map in the Sydney area, this environment requires an exact survey of the circumstances evoking the data properties at the time of collection. Even though tagging systems are a quickly evolving field, not all of the users have yet had the opportunity to accurately geotag their photos in a manual way, as the map and satellite data available on the Flickr interface is still badly resolved in many places of the world. Furthermore, user-generated data on Flickr does not cover all places well enough to derive useful information about locations at the sub-city level. The data is obviously not very dense in minor cities and less popular places. What has been called 'critical mass' by Weiss (2005) is not reached regarding the extraction of common-sense spatial knowledge and is largely missing for place tags in Zurich for instance. In order to estimate the future capabilities of Flickr data, the evolution of content which has been tagged within the cities of London and Sheffield since August 2006 was analysed. As shown in Figure 6.1, the database for both cities increases on a

linear scale, but at a very different absolute and also relative rate. For less popular places, such as Sheffield, it will obviously take a very long time to even come close to the level of coverage attained for major places in the very beginning of geotagging.

Despite the restrictions discussed here, the conclusions about the usage of terminology and the digital footprints generated for places of major cities such as London, Chicago, and Seattle show that the metadata produced by non-expert users holds valuable information. Users' tagging behaviour and hence the quality of user-contributed metadata seems adequate in the context of deriving knowledge about vernacular regions even at the neighbourhood level. If the bias introduced by single users is avoided, the results of this study support authors such as Mathes (2004), Guy and Tonkin (2006), and Rattenbury et al. (2007), who have pointed out the information potential of online tagging systems.

6.3 Approximation of footprints

The third research question is focussed on another problem inherent in the process of gaining knowledge about ill-defined and cognitive regions; the derivation of footprints from potential candidate points, which has been formulated as follows.

How can digital footprints of vernacular regions be modelled from georeferenced tags?

From the large body of work dealing with the generation of both fuzzy as well as crisp representations of vague regions (Alani et al., 2001; Purves et al., 2005; Schockaert and Cock, 2007; Twaroch et al., 2008) the approach of using KDE has been adopted within the scope of this work. A data-driven technique of bandwidth determination and threshold delineation was adopted to represent the probabilities of region membership not just for single places, but for a wide range of geographic terms and neighbourhood names. The evaluation of the method was carried out by means of public parks; spatial entities that can be considered as well-defined regardless the mode of observation and regionalisation. The evaluation approach is considered useful for further benchmarking of techniques to derive footprints from user-generated data.

Unsurprisingly, the performance of the approach was found to be dependent on the data that could be mined from Flickr. It worked more reliably for non-ambiguous, specific place names, which are more common than generic place tags and typically exhibited a rather correlated and single-peaked point pattern. The footprints derived from popular georeferenced tags showed an amazingly high agreement with official and conventional definitions. Where this was not the case, the estimation of density surfaces and their thresholding at the 90% contour line allowed for an assessment of reasons leading to the deviation. For some places, such as West Seattle, it could be shown that the common understanding differs from the semi-official definition, while others, such as the footprint of downtown Seattle, are obviously distorted by accumulations of outliers at scenic locations. For some places, for instance downtown Chicago, the footprints provide a means to establish where a non-defined place is commonly supposed to be located.

Despite the overall satisfying performance, some drawbacks of the approach were identified in the course of this project. The slight overestimation of highly clustered point patterns could possibly be diminished by the removal of the worst geometric outliers before the computation of the bandwidth parameter h_{ref} which results from the standard distribution. The structure of multiple-peaked data on the other hand will never be represented by the applied technique for bandwidth determination, as obvious from the distribution of centre in London. Furthermore, density estimation in general is susceptible to accumulations of outliers as well as to internal clusters, even if considering just one point per coordinate location. Therefore, not only the data, but also the resulting footprints are distorted towards scenic lookouts, landmarks, tourist attractions, the coastline (Seattle), or the lakeside (Zurich). As found by Henrich and Lüdecke (2008), KDE does not work well with regard to narrow and elongated features, but produces round shapes compared to the underlying data. Other typical properties of neighbourhoods, such as the abrupt change in the point pattern at clear-cut, bona fide boundaries constituted by rivers and street intersections are disregarded by the representations.

The computation of volume contour lines are an objective means to deal with possibly erroneous data. If the generation of crisp footprints in the form of polygons is required, the automatic choice of a threshold from this approach would be straightforward by cutting the surface at a given contour line. From the observations made in this project, the 80% and the 40% line could be considered as a preliminary suggestion of universal thresholds, representing a narrow and a wider approximation of the regions in the form of detailed polygons. This representation would account for the typical configuration of vague regions, which are said to be constituted by a core and a zone of transition (Murphy, 1972; Montello, 2003) and roughly correspond to the 'egg-yolk' model suggested by Cohn and Gotts (1996). A more data-driven rule for the choice of a threshold contour would require further empirical investigation focussed on this problem, but it might be obtained by some function of the number of reference points and their range of dispersion, represented by h_{ref}. The nature of the yielded representations is only considered limitedly useful for the automated generation of precise footprints as typically used in gazetteer services. The digital footprints stored in advanced gazetteers are, due to performance and storage capacity, mostly represented in the form of bounding boxes or simple polygons (Hill, 2006). To automatically derive footprints from Flickr data for the population of gazetteers, a more geometric method, such as a convex hull or even better a concave bounding hull representing the main point cluster, might be more suitable and useful.

Chapter 7

Conclusion

7.1 Accomplishments

To summarise, the contributions of this thesis are:

- A recapitulation of the nature of vague geographic regions and the implications on the formalisation and modelling of such regions within the framework of GIScience. A review of previous attempts to gain information about vague places motivated by problems of information retrieval as well as by other fields of the geographic or related disciplines. An overview of considerations and research regarding the emergence of user-contributed tagging systems on the web.

- The successful extraction of a large amount of spatially relevant data from the Flickr database, for which previously found patterns in user-contributed content, such as the power law distribution of used tags, the proportion of unique tags, the average number of tags per photo, and the strong bias by single users, could be confirmed. An assessment of the reasonable quality of textual and formal place tags in user-generated metadata.

- The adoption of an explorative approach to extract information about common-sense geographical reasoning and knowledge from a large amount of user-contributed metadata, guided by data-driven as well as theory-driven techniques.

- An in-depth analysis of the nature of intuitive place indications in natural language tags. The establishment of global and local patterns of place tags in the German and English language use. An investigation of the possible interrelation between formal geotags and semantic place references.

- An investigation of the conditions required to extract distributed knowledge and shared cognition from large samples of user-contributed metadata.

- The development of an objective approach to derive vague footprints at the sub-city level of granularity from georeferenced Flickr tags based on KDE, including an automatic way to deal with potential outliers. A proposal for a repeatable methodology to derive sharp boundaries from the cognitive representations.

7.2 Findings

In the framework of this project, the creation of ad-hoc keywords to categorise georeferenced resources in online tagging systems was considered as a proxy of how people intuitively refer to location. The quantitative evaluation of tag usage in different cities revealed that a large proportion of the top-ranked tags in georeferenced Flickr samples correspond to place names. While the exact values are biased by the chosen frame of reference, there are clear indications that the city toponym is prevalent over lower-level and higher-level place indications. Regardless of cultural and linguistic backgrounds, the official toponym is by far the most essential frame of reference and could be said to be the basic level of geographic tags. Except for downtown in the US and CBD in Australia, generic city core terms are not frequently employed as tags and tend to exhibit bias with regard to user ubiquity, especially in minor places such as Zurich and Sheffield, as well as in London. Hence, it is difficult to establish significant patterns for these kinds of terms. Instead, it can be said that the usage of generic terminology is strongly place-dependent. Specific place names of districts and neighbourhoods occur more prominently in the taglists, particularly if associated with larger cities. Apparently, people think in terms of concrete places and their names, rather than functional concepts when describing space in the annotation of online resources, even though the latter could be considered more useful with respect to the idea of sharing and providing information for a wide range of users perhaps not familiar with a particular place.

Formal geotags related to neighbourhood labels are highly spatially auto-correlated, despite the complex nature of cognitive processes and the distributed and uncoordinated process of tagging. The results of this study suggest that the average user has a distinct idea of specific places, their location, and their extent. The generation of spatial footprints from Flickr data is straightforward and the findings reveal that the users' overall attitude towards the creation of metadata meets the requirements for the generation of footprints for practical purposes at the sub-city level of granularity. However, the performance of the approach is directly linked to the popularity of a place tag. Furthermore, it does not yet work reliably for areas which are less well represented on Flickr, such as minor or less popular places. Due to the essential nature of photography, the data is also highly susceptible to internal clustering. The representations of cognitive regions using KDE and thresholding the surfaces at the 90% volume contour is suitable for the investigation of both the vague aspect of urban places and shortcomings in the Flickr data. Other properties of the data structure, such as the delimitation of neighbourhoods at clear-cut, bona fide boundaries, are not captured by the representations. Regarding the needs of gazetteer services, geographically related information services, and the providers of map products, the presented approach of footprint approximation is much less labour intensive

than human subject tests and more scalable than approaches based on an experimental choice of parameters. For the generation of manageable representations, a more geometric method, such as the derivation of convex or concave hulls, should probably be considered.

The findings confirm that the user-generated metadata in the Flickr database is an immense source of spatially relevant and valuable empirical data. At high levels of frequency and popularity, the chaotic nature of tags converges towards the emergence of shared conventions. At the same time, the layout of the system allows the single user to produce a large amount of data effortlessly. The constraints resulting from this fact need to be carefully taken into consideration to draw objective conclusions. In view of the evolution of the amount of content being tagged in different places, it is questionable whether Flickr will soon cover minor cities well enough for the purposes here discussed. Given the 'critical mass', the abundance and quality of formal and textual place references in the Flickr data support its suitability for the extraction of common-sense spatial knowledge. This could be relevant in scientific disciplines beyond GIR and might be used for a wide range of purposes. The greatest advantage of the information source in this context is seen in its topicality and adaptiveness, as places within urban environments are known to emerge and evolve, and the conceptions of neighbourhood boundaries tend to change constantly. The Flickr data is considered as highly beneficial for capturing current, local understandings of places in many parts of the world and might constitute the basis of much future investigation in this context.

7.3 Future directions and suggestions

By means of the large-scale quantitative analysis carried out within the scope of this project, the city level has been established as the granularity level people most intuitively think of when assigning locational information. Furthermore, the nature and shortcomings of user-generated geotags have been investigated. Regarding problems of system design, human computer interaction as well as cognitive science, it would be very interesting to better understand the conditions under which people create formal and semantic location information of a certain kind. It remains to be verified how the creation of spatially related content by ordinary users depends on their orientation ability, their technical expertise, or their familiarity with a place. The findings of cognitive science imply that there are, for instance, considerable differences in the spatial reasoning of people being familiar or new to a place. By categorising users into residents and visitors, as accomplished by Girardin et al. (2008), it could be analysed whether and to what extent their employment of formal and semantic place tags varies in terms of accuracy and spatial distribution. More seminal insights into users' intentions and cognitive choices upon assigning particular place tags might be attained by checking annotations against the visual content of photographs or by directly interviewing users.

In view of the availability of an immense multilingual database, an additional research avenue could be motivated from a language geographic or linguistic point of view. Within this project, the location of residence indicated by those people who had used a particular

generic place tag was analysed at a global level, but not within the extent of a specific city. The cultural and linguistic background of the users contributing the downtown tags within the UK, for instance, was not verified and could only be guessed. Therefore we do not know, if it is actually the terminology which is expanding to members of other cultures or if it is rather the people themselves diffusing.

With regard to the objective of the automated derivation of crisp digital footprints, the approach here employed requires further empirical investigation. The rule for the choice of a threshold contour line or the applicability of a more geometric method to derive footprints could be further explored. Many useful approaches have been presented to generate footprints of geographic regions, but it is to be established what kind of representations are suitable for which purpose (e.g. Davies et al., 2008). Another intriguing question raised in the course of this project is the definition of 'vernacular' versus 'official' regions. In practice, the terms are rather context-dependent than opposed. To date, the approaches presented to derive digital footprints have usually been applied to single, specific regions, which were well known to the respective authors and corresponded rather to the vernacular or to the official category. The automated identification of vernacular place names and a "method to measure the degree of vernacularity" (Twaroch et al., 2008: 64) has not been addressed to date. It is not even clear which type of spatial entities apply to the concept of place in this context (Davies et al., 2008). With regard to the domains beyond GIR interested in the interpretation of place names, a major focus for future work seems therefore the establishment of (automated) approaches to obtain large-scale collections of commonly used place names and their spatial relationships. The availability of such information is considered as one of the main benefits of the geocoded Flickr content, as respective knowledge is typically dispersed and local and not held by a single institution or a group of experts. For instance, the application of the method presented by Rattenbury et al. (2007) to automatically extract tags representing locations at different levels of scale could be revised in order to generate lists of candidate names with associated point data.

The availability of extensive collection of data from Flickr though depends on the goodwill of the operator and might furthermore not be sufficient for the areas less well covered by the platform. Therefore, it should also be explored how user-generated locational data can be combined and integrated with data from other sources. Geotagged hypermedia provides a platform for innovate approaches in the context of common-sense perception and knowledge of geographic space. Within the framework of this thesis it has been disclosed why and how GIScience can benefit from such information. Future work should be directed towards the formalisation of methods and concepts to deal with this new kind of geographic information and finally towards the ability to link traditional and new geodata together.

Bibliography

Ahern, S., Naaman, M., Nair, R., Yang, J. H.-I. (2007): World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In: JCDL '07: Proceedings of the 2007 conference on digital libraries, 1–10, ACM, New York, NY, US.

Alani, H., Jones, C., Tudhope, D. (2001): Voronoi-Based Region Approximation for Geographical Information Retrieval with Gazetteers. International Journal of Geographical Information Science, 15(4), 287–306.

Ames, M., Naaman, M. (2007): Why we tag: motivations for annotation in mobile and online media. In: CHI '07: Proceedings of the SIGCHI conference on human factors in computing systems, 971–980, ACM Press, New York, US.

Arampatzis, A., van Kreveld, M., Reinbacher, I., Jones, C. B., Vaid, S., Clough, P., Joho, H., Sanderson, M. (2006): Web-based delineation of imprecise regions. Computers, Environment and Urban Systems, 30(4), 436–459.

Boll, S., Jones, C., Kansa, E., Kishor, P., Naaman, M., Purves, R., Scharl, A., Wilde, E. (Eds.) (2008): Location and the Web, LocWeb 2008, WWW 2008 Conference, Beijing, China.

Burrough, P. (1996): Natural Objects with Indeterminate Boundaries. In: Burrough, P., Frank, A. (Eds.), Geographic Objects with Indeterminate Boundaries, Gisdata 2, 3–28, Taylor & Francis Ltd, London, UK.

Burrough, P., Frank, A. (1996): Geographic Objects with Indeterminate Boundaries. Gisdata2, Taylor & Francis Ltd, London, UK.

Burrough, P. A., McDonnell, R. A. (1998): Principles of Geographical Information Systems. Spatial Information Systems and Geostatistics, Oxford University Press Inc., New York, NY, US.

Byrkit, J. W. (1992): Land, Sky, and People: The Southwest Defined. University of Arizona Press, Tuscon, Arizona, US.

Campari, I. (1996): Uncertain Boundaries in Urban Space. In: Burrough, P., Frank, A. (Eds.), Geographic Objects with Indeterminate Boundaries, Gisdata 2, 57–69, Taylor & Francis Ltd, London, UK.

Catt, D. (2008): Going places on flickr: The significance of geographical information in photos. Presentation at Where 2.0, Burlingame, CA, US.

Caves, R. W. (Ed.) (2005): Encyclopedia of the city. Routledge, Taylor & Francis Group, Oxon, England.

Cohn, A., Gotts, N. (1996): The 'Egg-Yolk' Representation of Regions with Indeterminate Boundaries. In: Burrough, P. A., Frank, A. U. (Eds.), Geographic Objects with Indeterminate Boundaries, Gisdata 2, 171–187, Taylor & Francis Ltd, London, UK.

Couclelis, H. (1996): Towards an Operational Typology of Geographic Entities with Ill-defined Boundaries. In: Burrough, P., Frank, A. (Eds.), Geographic Objects with Indeterminate Boundaries, Gisdata 2, 45–55, Taylor & Francis Ltd, London, UK.

Couclelis, H. (1998): Aristotelian Spatial Dynamics in the Age of Geographic Information Systems. In: Egenhofer, M. J., Golledge, R. (Eds.), Spatial and Temporal Reasoning in Geographic Information Systems, Spatial Information Series, 109–118, Oxford University Press Inc., New York, NY, US.

Couclelis, H. (2003): The Certainty of Uncertainty: GIS and the Limits of Geographic Knowledge. Transactions in GIS, 7(2), 165–175.

Davies, C., Holt, I., Green, J., Harding, J., Diamond, L. (2008): User Need and the Implications for Modelling Place. In: Winter, S., Kuhn, W., Krüger, A. (Eds.), International Workshop on Computational Models of Place, PLACE'08, 1–14, GIScience'08, The University of Melbourne, Park City, Utah, US.

de Smith, M., Goodchild, M., Longley, P. (2008): Geospatial Analysis - a comprehensive guide. URL `http://www.spatialanalysisonline.com/output/html/Pointdensity.html`.

Dubinko, M., Kumar, R., Magnani, J., Novak, J., Rghavan, P., Tomkinsk, A. (2006): Visualizing Tags over Time. URL `http://www2006.org/programme/item.php?id=25`.

Egenhofer, J., Golledge, R. (1998): Spatial and Temporal Reasoning in Geographic Information Systems. Spatial Information Series, Oxford University Press Inc., New York, US.

Egenhofer, J., Mark, D. (1995): Naïve Geography. In: Frank, A., Kuhn, W. (Eds.), Naïve Geography COSIT '95, vol. 988, 1–15, Lecture Notes in Computer Science, Springer-Verlag, Berlin / Heidelberg, Germany.

Erle, S., Gibson, R., Walsh, J. (2005): Mapping Hacks. Hacks Series, O'Reilly Media, Inc., Sebastopol, CA, US.

Evans, A. (2004): Oop'Narf and Up The Junction: Capturing the Vernacular. Tech. rep., Nottingham University, URL `www.geog.leeds.ac.uk/presentations/04-5/04-5.ppt`.

Ferrari, G. (1996): Boundaries, Concepts, Language. In: Burrough, P., Frank, A. (Eds.), Geographic Objects with Indeterminate Boundaries, Gisdata 2, 99–108, Taylor & Francis Ltd, London, UK.

Fisher, P. (1996): Boolean and Fuzzy Regions. In: Burrough, P., Frank, A. (Eds.), Geographic Objects with Indeterminate Boundaries, Gisdata 2, 87–94, Taylor & Francis Ltd., London, UK.

Fisher, P. (1999): Models of uncertainty in spatial data. In: Longley, P. A., Goodchild, M. F., Maguire, D. J., Rhind, D. W. (Eds.), Geographical Information Systems Principles and Technical Issues, vol. 1, 191–205, John Wiley & Sons Ltd, New York, NY, US.

Fogelson, R. M. (2001): Downtown. Its Rise and Fall, 1880-1950. R.R. Donnelley & Sons, Harrisonburg, Virginia, US.

Frank, A. (1996): The Prevalence of Objects with Sharp Boundaries in GIS. In: Burrough, P., Frank, A. (Eds.), Geographic Objects with Indeterminate Boundaries, Gisdata 2, 29–40, Taylor & Francis Ltd, London, UK.

Gaebe, W. (2004): Urbane Räume. Eugen Ulmer GmbH & co., Stuttgart, Germany.

Gale, N., Golledge, R. (1982): On the subjective partitioning of space. Annals of the Association of American Geographers, 72(1), 60–67.

Gan, Q., Attenberg, J., Markowetz, A., Suel, T. (2008): Analysis of Geographic Queries in a Search Engine Log. In: Proceedings of the First International Workshop on Location and the Web, 49–56, LocWeb 2008, 17th International World Wide Web Conference, Beijing, China.

Girardin, F., Blat, J. (2007): Place this Photo on a Map: A Study of Explicit Disclosure of Location Information. 9th International Conference on Ubiquitous Computing (UbiComp 2007), Innsbruck, Austria, URL `www.girardin.org/fabien/publications/girardin_ubicomp2007_lbr.pdf`.

Girardin, F., Blat, J., Calabrese, F., Dal Fiore, F., Ratti, C. (2008): Digital Footprinting: Uncovering Tourists with User-generated Content. IEEE Pervasive Computing, 7(4), 36–43.

Golder, S., Huberman, B. A. (2005): The Structure of Collaborative Tagging Systems. URL `http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0508082`.

Goodchild, M., Montello, D., Fohl, P., Gottsegen, J. (1998): Fuzzy spatial queries in digital spatial data libraries. Fuzzy Systems Proceedings, 1998. IEEE World Congress on Computational Intelligence, 1, 205–210.

Grothe, C., Schaab, J. (2008): An Evaluation of Kernel Density Estimation and Support Vector Machines for Automated Generation of Footprints for Imprecise Regions from Geotags. In: Winter, S., Kuhn, W., Krüger, A. (Eds.), International Workshop on Computational Models of Place, PLACE'08, 15–28, GIScience'08, The University of Melbourne, Park City, Utah, US.

Guy, M., Tonkin, E. (2006): Folksonomies: Tidying up Tags? URL `http://www.dlib.org/dlib/january06/guy/01guy.html`.

Hassan-Montero, Y., Herrero-Solana, V. (2006): Improving Tag-Clouds as Visual Information Retrieval Interfaces. In: InScit2006: International Conference on Multidisciplinary Information Sciences and Technologies, Mérida, Spain.

Hastings, J. (2008): Automated conflation of digital gazetteer data. International Journal of Geographical Information Science, 22(10), 1109–1127.

Heineberg, H. (2000): Grundriss Allgemeine Geographie: Stadtgeographie. Verlag Ferdinand Schöningh, Paderborn, Germany.

Henrich, A., Lüdecke, V. (2008): Determining Geographic Representations for Arbitrary Concepts at Query Time. In: Proceedings of the First International Workshop on Location and the Web, 17–24, LocWeb 2008, 17th International World Wide Web Conference, Beijing, China.

Hill, L. (2006): Georeferencing: the geographic associations of information. Digital Libraries and Electronic Publishing, The MIT Press, Cambridge, Massachusetts, US.

Hill, L., Frew, J., Zheng, Q. (1999): Geographic Names. The Implementation of a Gazetteer in a Georeferenced Digital Library. D-Lib Magazine, 5(1).

Hirtle, S. C. (2003): Neighborhoods and landmarks. In: Duckham, M., Goodchild, M. F., Worboys, M. F. (Eds.), Foundations of Geographic Information Science, Taylor & Francis Ltd., London, UK.

Hofmeister, B. (1996): Die Stadtstruktur. Ihre Ausprägungen in den verschiedenen Kulturräumen der Erde. 132, Wissenschaftliche Buchgesellschaft, Darmstadt, Germany.

Jones, C., Purves, R., Clough, P., Joho, H. (2008): Modelling Vague Places with Knowledge from the Web. International Journal of Geographical Information Science, 22(10), 1045–1065.

Juchelka, R. (2001): Zentral-Zentrum-Zentrierung. Eine theoretisch-terminologische Diskussion zu traditionellen Begriffen der Geographie, ihren aktuellen Adaptionen und planungspraktischen Anwendungen. In: Wohlschlägl, H. (Ed.), Geographischer Jahresbericht aus Österreich, vol. LVIII, 67–81, Ferdinand Berger & Söhne GmbH, Wien, Austria.

Kennedy, L., Chang, S.-F., Kozintsev, I. (2006): To search or to label?: predicting the performance of search-based automatic image classifiers. In: Proceedings of the 8th ACM international workshop on Multimedia information retrieval, p. 249–258.

Kennedy, L., Naaman, M., Ahern, S., Nair, R., Rattenbury, T. (2007): How flickr helps us make sense of the world: context and content in community-contributed media collections. In: MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia, 631–640, ACM, New York, NY, US.

King, L., Golledge, R. (1978): Cities, Space, and Behavior. Prentice-Hall, Inc., Englewood Cliffs, US.

Kitchen, R., Blades, M. (2002): The Cognition of Geographic Space. I.B. Tauris & Co Ltd., London, UK.

Knox, P., Pinch, S. (2000): Urban Social Geography: An Introduction. Pearson Education, Prentice-Hall, Inc., Essex, England, UK.

Lakoff, G. (1987): Women, Fire, and Dangerous Things. The University of Chicago Press, Chicago, Il, US.

Lam, C., Wilson, J., Holmes-Wong, D. (2002): Building a Neighborhood-Specific Gazetteer for a Digital Archive. URL `http://gis.esri.com/library/userconf/proc02/pap0300/p0300.htm`.

Larson, R. R. (1995): Geographic information retrieval and spatial browsing. In: Smith, L., Gluck, M. (Eds.), Geographic Information Systems and Libraries: Patrons, Maps and Spatial Information, 81–124, NN, University of Illinois at Urbana-Champaign, US.

Law, C. M. (1988): The uncertain future of the urban core. Routledge, London, UK.

Lerman, K., Jones, L. (2006): Social Browsing on Flickr. URL `http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0612047`.

Llyod, W. (1976): Landscape Imagery in the Urban Novel: A Source of Geographic Evidence. In: Moore, G., Golledge, R. (Eds.), Environmental Knowing, vol. 23, 279–285, Dowden, Hutchingon & Ross, Inc., Stroudsburg, Pennsylvania, US.

Lynch, K. (1960): The Image of the City. M.I.T. Press & Harvard University Press, Cambridge, Massachusetts, UK.

Macrgegor, G., McCulloch, E. (2006): Collaborative tagging as a knowledge organisation and resource discovery tool. URL `www.emeraldinsight.com/0024-2435.htm`.

Mansbridge, L. (2005): Perceptions of Imprecise Regions in Relation to Geographical Information Retrieval. Msc thesis, University of Sheffield.

Mark, D., Freska, C., Hirtle, C., Lloyd, R., Tversky, B. (1999): Cognitive models of geographic space. International Journal of Geographical Information Science, (13), 747–774.

Mark, D. M., Turk, A. G. (2003): Landscape Categories in Yindjibarndi: Ontology, Environment, and Language. In: Kuhn, W., Worboys, M. F., Timpf, S. (Eds.), Spatial Information Theory: Foundations of Geographic Information Science, Lecture Notes in Computer Science, 28–45, International Conference, COSIT 2003, Springer-Verlag, Berlin/Heidelberg, Germany, Kartause, Ittingen, Switzerland.

Marlow, C., Naaman, M., Boyd, D., Davis, M. (2006): HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia, 31–40, ACM, New York, NY, US.

Mathes, A. (2004): Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Computer Mediated Communication - LIS590CMC, URL `http://blog.namics.com/2005/Folksonomies_Cooperative_Classification.pdf`.

McGranaghan, M. (1990): Matching Representations of Geographic Locations. In: Cognitive and Linguistic Aspects of Geographic Space, vol. 90, 32–48, National Center of Geographic Information & Analysis NCGIA, University of Maine, ME, US.

McNamara, T. P. (1986): Mental representations in spatial relations. Cognitive Psychology, (18), 87–121.

Montello, D. R. (1995): How significant are cultural differences in spatial cognition? In: Frank, A. U., Kuhn, W. (Eds.), Spatial information theory: A theoretical basis for GIS, *Lecture Notes in Computer Science*, vol. 988, 485–500, Springer-Verlag, Berlin / Heidelberg, Germany.

Montello, D. R. (1998): A New Framework for Understanding the Acquisition of Spatial Knowledge in Large-Scale Environments. In: Egenhofer, M. J., Golledge, R. (Eds.), Spatial and Temporal Reasoning in Geographic Information Systems, Spatial Information Series, 143–154, Oxford University Press Inc., New York, US.

Montello, D. R. (2001): Spatial Cognition. In: Smelser, N. J., Baltes, P. B. (Eds.), International Encyclopedia of the Social & Behavioral Sciences, 14771–14775, Pergamon Press, Oxford, UK.

Montello, D. R. (2003): Regions in geography: Process and content. In: Duckham, M., Goodchild, M. F., Worboys, M. F. (Eds.), Foundations of Geographic Information Science, 173–189, Taylor & Francis, London.

Montello, D. R., Freundschuh, S. (1995): Sources of spatial knowledge and their implications for GIS: An introduction. Geographical Systems, 2, 169–176.

Montello, D. R., Freundschuh, S. (2005): Cognition of Geographic Information. In: McMaster, R. B., Usery, E. L. (Eds.), A Research Agenda for Geographic Information Science, 61–91, CRC Press., Boca Raton, FL, US.

Montello, D. R., Goodchild, M. F., Gottsegen, J., Fohl, P. (2003): Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. Spatial Cognition & Computation, 3(2-3), 185–204.

Murphy, R. E. (1972): The Central Business District. Aldine Atherton, Inc., Chicago, Il, US.

O'Sullivan, D., Unwin, D. J. (2003): Geographic Information Analysis. John Wiley & Sons, Inc., Hoboken, New Jersey, US.

Purves, R., Clough, P., Joho, H. (2005): Identifiying imprecise regions for geographic information retrieval using the web. In: Billen, R., Drummond, J., Forrest, D., Joao, E. (Eds.), Proceedings of the GIS RESEARCH UK 13th Annual Conference, 313–318, Glasgow, UK.

Purves, R., Jones, C. (2006): Geographic Information Retrieval (GIR). Computers, Environment and Urban Systems, 30(4), 375–377.

Purves, R., Jones, C. (2007): The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. International Journal of Geographical Information Science, 21(7), 717–745.

Rapoport, A. (1976): Environmental Cognition in Cross-Cultural Perspective. In: Moore, G., Golledge, R. (Eds.), Environmental Knowing, *Community Development Series*, vol. 23, 220–234, Dowden, Hutchingon & Ross, Inc., Stroudsburg, Pennsylvania, US.

Rattenbury, T., Good, N., Naaman, M. (2007): Towards automatic extraction of event and place semantics from flickr tags. In: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 103–110, ACM, New York, NY, US.

Rodgers, A., Carr, A. (1998): HRE: The Home Range Extension for ArcView. User's Manual. Centre for Northern Forest Ecosystem Research, Ontario, URL `http://blue.lakeheadu.ca/hre/`.

Rosch, E. (1978): Principles of categorization. In: Rosch, E., Lloyd, B. (Eds.), Cognition and Categorization, 27–48, NN, Erlbaum.

Sanderson, M., Kohler, J. (2004): Analyzing geographic queries. URL `http://dis.shef.ac.uk/mark/publications/my_papers/GeoQueryAnalysis2004.pdf`.

Schmitz, P. (2006): Inducing ontology from Flickr tags. In: Proc. of the Collaborative Web Tagging Workshop (WWW '06), URL `http://www.rawsugar.com/www2006/22.pdf`.

Schockaert, S., Cock, M. D. (2007): Neighborhood restrictions in geographic IR. In: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, 167–174, ACM, New York, NY, US.

Schönauer, I. (2007): Bevölkerung Stadt Zürich. Stadt Zürich, Präsidialdepartement Statistik Stadt Zürich, 4 edn.

Sen, S. (2008): Characterizing Places in geospatial ontologies: Specifying partial knowldege about their use. In: Winter, S., Kuhn, W., Krüger, A. (Eds.), International Workshop on Computational Models of Place, PLACE'08, 29–44, GIScience'08, The University of Melbourne, Park City, Utah, US.

Shirky, C. (2005): Ontology is overrated: Categories, Links, and Tags. URL `http://www.shirky.com/writings/ontology_overrated.html`.

Smith, B. (1995): On drawing lines on a map. In: Frank, A., Kuhn, W. (Eds.), COSIT'95, Lecture Notes in Computer Science 988, 475–484, Springer-Verlag, Berlin / Heidelberg, Germany.

Smith, B., Mark, D. M. (2001): Geographical Categories: An Ontological Investigation. International Journal of Geographical Information Science, 15(7), 591–612.

Smith, B., Varzi, A. C. (2000): Fiat and Bona Fide Boundaries. Philosphy and Phenomenological Research, 60(2), 401–420.

Smith, G. (2004): Folksonomy: Social Classification. URL `http://atomiq.org/archives/2004/08/folksonomy_social_classification.html`.

Steels, L. (2006): Collaborative tagging as distributed cognition. Pragmatics & Cognition, 14(2), 287–292.

Sturtz, D. N. (2004): Communal Categorization: The Folksonomy. INFO622: Content Representation, URL `www.davidsturtz.com/drexel/622/sturtz-folksonomy.pdf`.

Talen, E. (1999): Constructing neighborhoods from the bottom up: the case for resident-generated GIS. Environment and Planning B: Planning and Design, 26, 533–554.

Tanasescu, V., Domigue, J. (2008): A Differential Notion of Place for Local Search. In: Proceedings of the First International Workshop on Location and the Web, 9–16, LocWeb 2008, 17th International World Wide Web Conference, Beijing, China.

Thurstain-Godwin, M., Unwin, D. J. (2000): Defining & delineating the central areas of towns for statistical monitoring using continuous surface representations. Tech. Rep. 18, CASA, Centre for advanced spatial analysis, London, URL `http://eprints.ucl.ac.uk/1363/`.

Twaroch, F. A., Jones, C. B., Abdemoty, A. I. (2008): Acquisition of a Vernacular Gazetteer from Web Sources. In: Proceedings of the First International Workshop on Location and the Web, 61–64, LocWeb 2008, 17th International World Wide Web Conference, Beijing, China.

Vogele, T., Schlieder, C., Visser, U. (2003): Intuitive modelling of place names for spatial information retrieval. In: Kuhn, W., Worboys, M. F., Timpf, S. (Eds.), Proceedings of COSIT'03, 239–52, Lecture Notes in Computer Science 2825, Springer-Verlag, Berlin / Heidelberg, Germany.

Waters, T., Evans, A. (2003): Tools for the web-based GIS mapping of "fuzzy" vernacular geography. In: Proceedings of the 7th International Conference on GeoComputation, Southampton, UK.

Weinberger, D. (2007): Tagging and Why It Matters. URL `http://cyber.law.harvard.edu/home/uploads/507/07-WhyTaggingMatters.pdf`.

Weiss, A. (2005): The power of collective intelligence. netWorker, 9(3), 16–23.

Winget, M. (2006): User-defined Classification on the Online Photo Sharing Site Flickr... or, how I Learned to Stop Worrying and Love the Million Typing Monkeys. In: Furner, J., Tennis, J. T. (Eds.), Advances in classification research, Vol. 17: Proceedings of the 17th ASIS&T SIG/CR Classification.

Zadeh, L. (1965): Fuzzy sets. Information and Control, (8), 338–353.

Zhang, V., Rey, B., Stipp, E., Jones, R. (2006): Geomodification in query rewriting. In: Proceedings of the 3. Workshop on Geographic Information Retrieval, GeoIR 2006, Seattle, WA, US.

Appendix A

Flickr data

A.1 Bounding box coordinates for spatial search

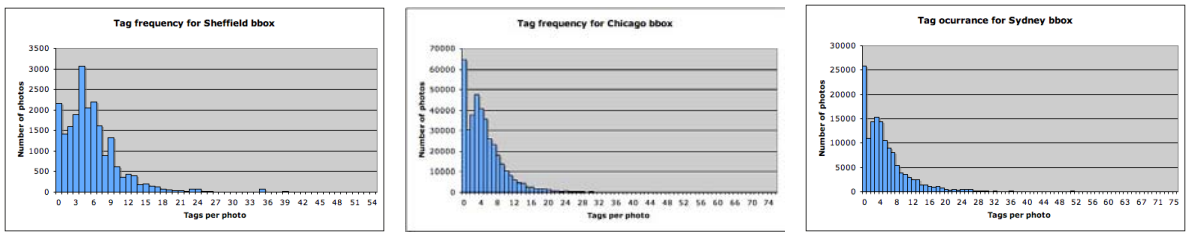
bbox	min_longitude	min_latitude	max_longitude	max_latitude
Zurich	8.45991944	47.3390138	8.5849000	47.41896944
London	-0.63152777	51.2364638	0.38765555	51.72694444
Sheffield	-1.531008333	53.3581722	-1.402986110	53.41547500
Chicago	-87.95913611	41.5990833	-87.45048055	42.106986111
Seattle	-122.4675166	47.51952777	-122.23838611	47.74945000
Sydney	151.12606944	-33.9365666	151.258225	-33.84986944

Table A.1: Bounding coordinates used for spatial search of different cities

A.2 Characteristics of Flickr data

A.2.1 Number of tags per item

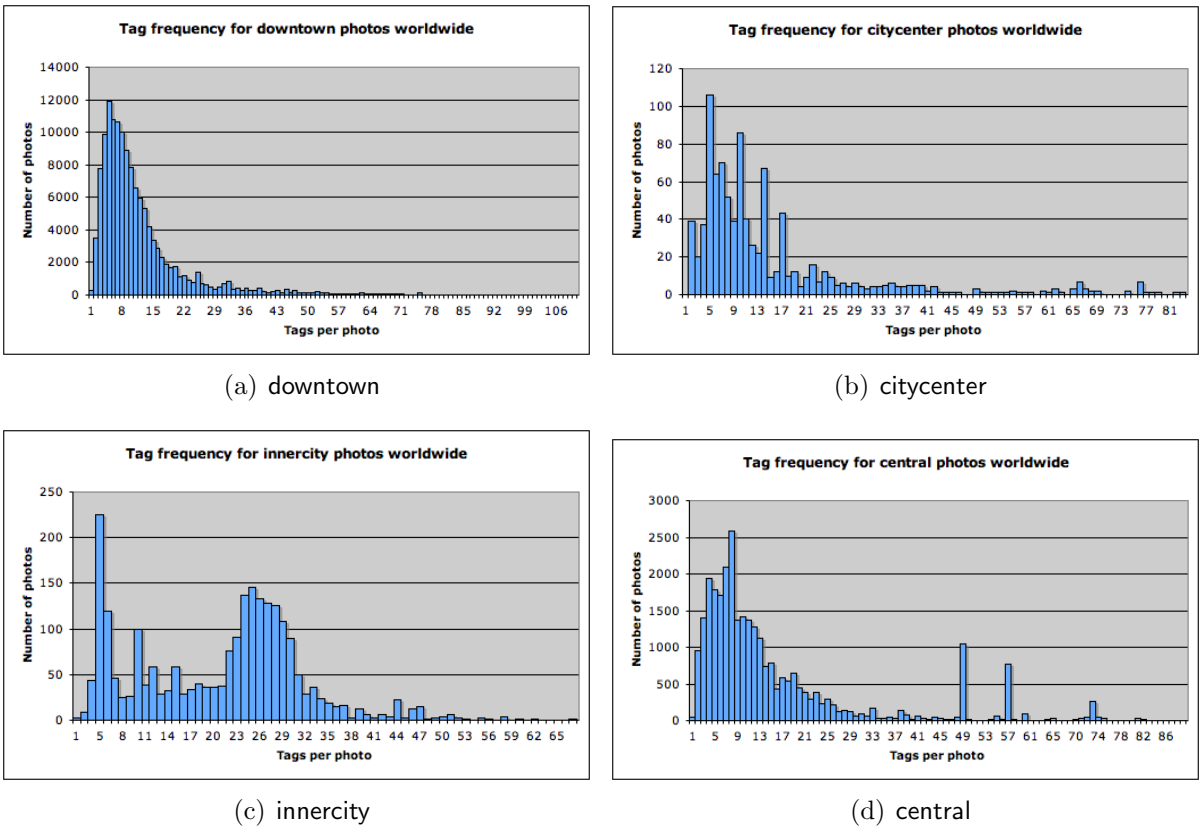
City-dataset



(a) Sheffield (b) Chicago bbox (c) Sydney

Figure A.1: Tag frequency within the bounding boxes of different cities

Global-dataset

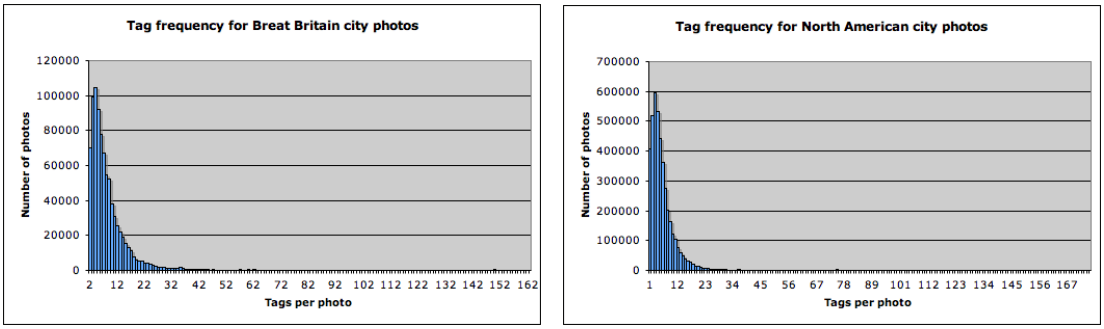


(a) downtown (b) citycenter

(c) innercity (d) central

Figure A.2: Tag frequency for different tags on the global level

Region-dataset

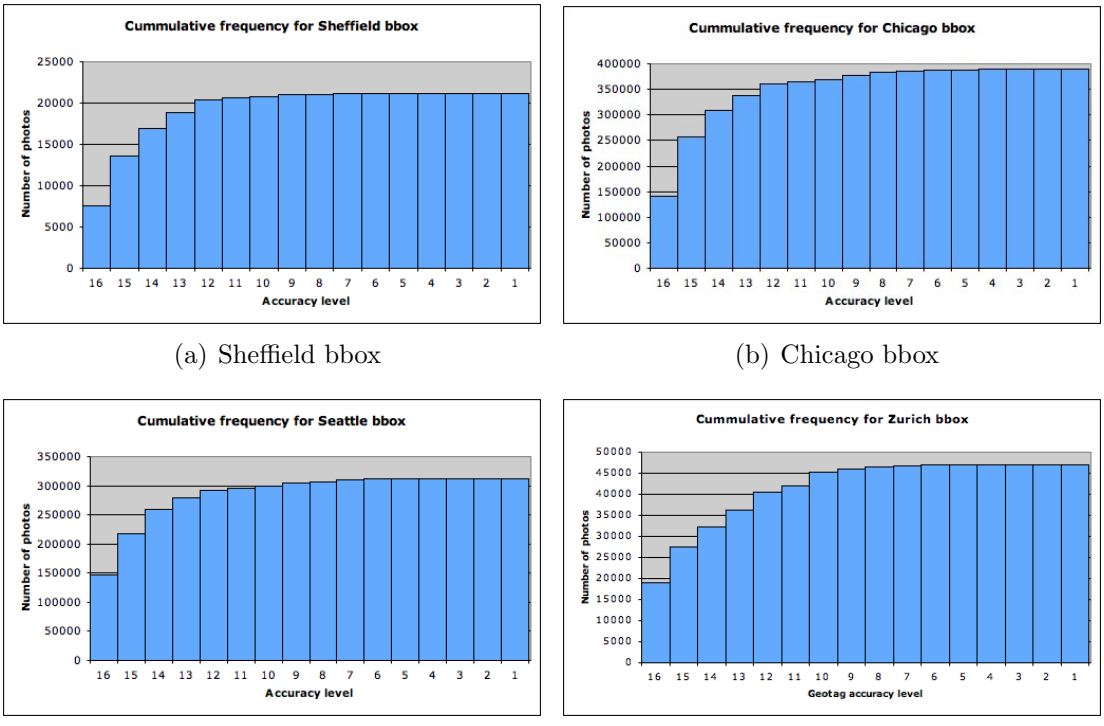


(a) GB cities (b) US cities

Figure A.3: Tag frequency for non-georeferenced items associated with a specific toponym tag

A.2.2 Geotag accuracy

City-dataset

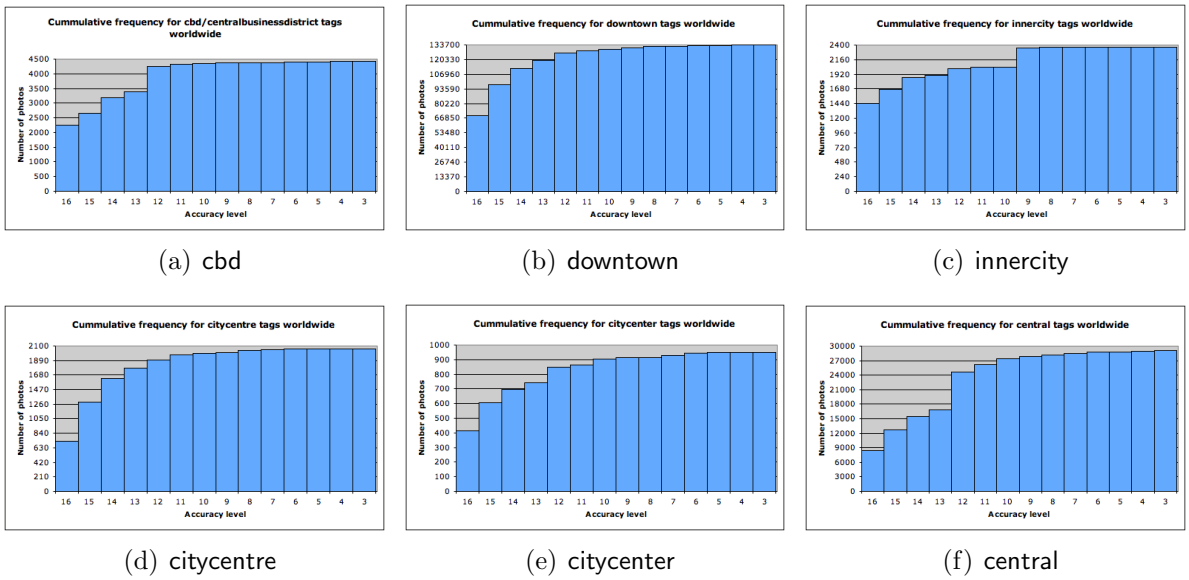


(a) Sheffield bbox (b) Chicago bbox

(c) Seattle bbox (d) Zurich bbox

Figure A.4: Cumulative frequency of geotag level for georeferenced data within different bounding boxes

Global-dataset



(a) cbd (b) downtown (c) innercity

(d) citycentre (e) citycenter (f) central

Figure A.5: Cumulative frequency of geotag level for georeferenced data sets associated with different city core tags

A.2.3 Spatial distribution of geotagged items

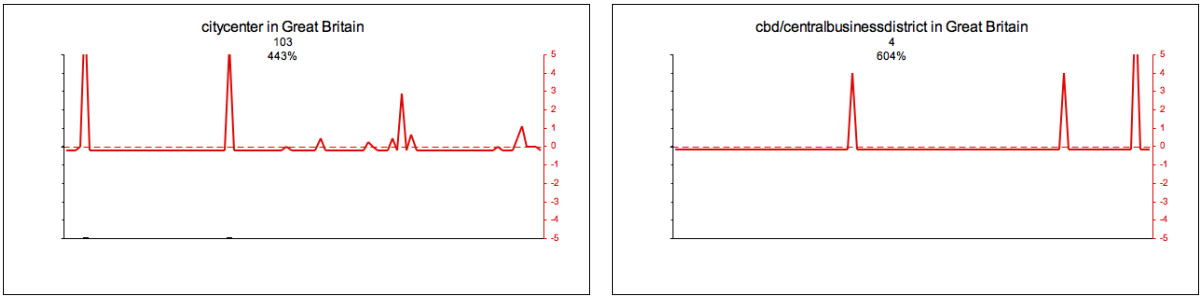
(a) Sheffield (b) Seattle (c) Chicago

Figure A.6: Spatial distribution of all georeferenced items within the bounding boxes of different cities

Appendix B

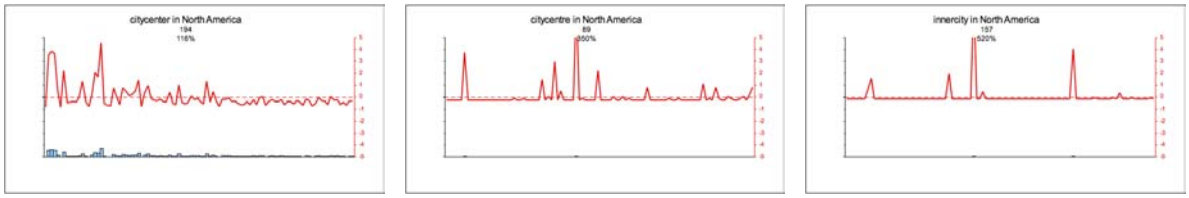
Data analysis

B.1 Tag profiles at regional level



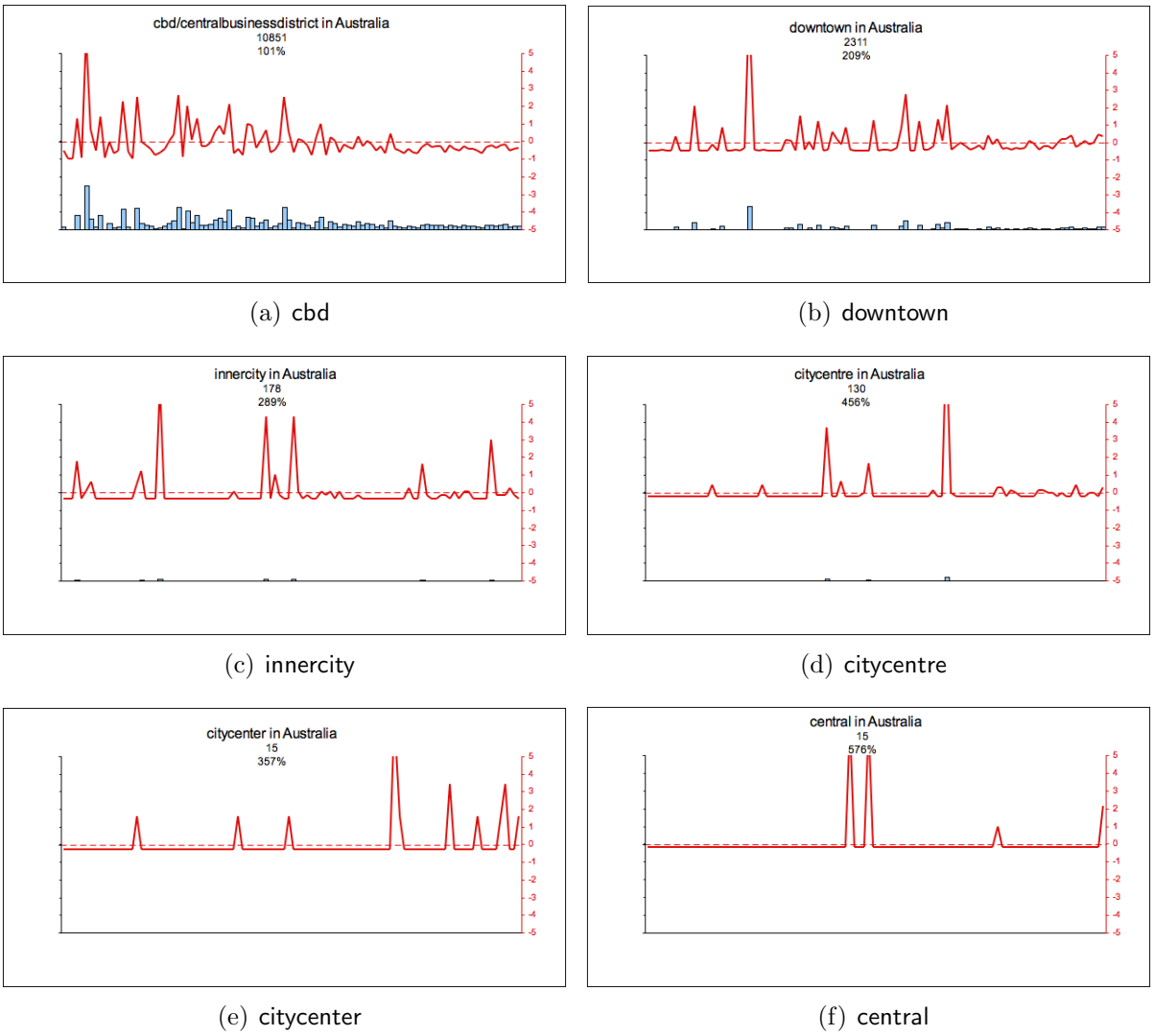
(a) citycenter (b) cbd

Figure B.1: Tag profiles for city core tags associated with different toponyms tags of British cities



(a) citycenter (b) citycentre (c) innercity

Figure B.2: Tag profiles for city core tags associated with different toponym tags of US cities



(a) cbd (b) downtown

(c) innercity (d) citycentre

(e) citycenter (f) central

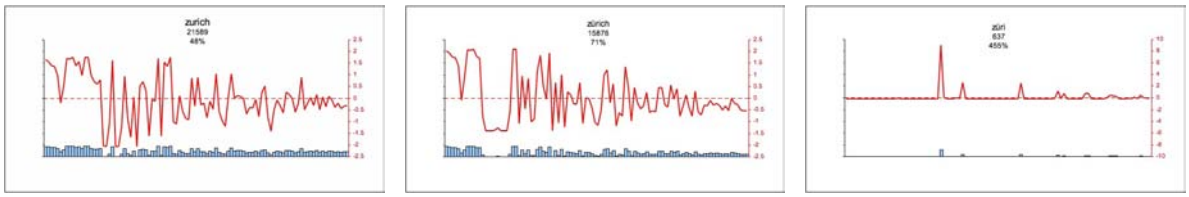
Figure B.3: Tag profiles for city core tags associated with Australian cities

B.2 Analysis at the city level

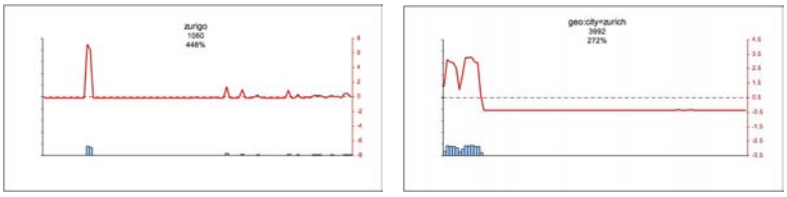
Zurich			
tag	**number**	**% of tags**	**C. of var.**
zurich	21'589	8.51%	48%
zürich	15'876	6.3%	71%
geo:city=zurich	3'992	1.6%	272%
zuerich	1'487	0.59%	277%
zurigo	1'060	0.42%	448%
züri	637	0.25%	545%
other		0.51%	
total		**18.1%**	
London			
tag	**number**	**% of tags**	**C. of var.**
london	539'175	16.40%	18%
londres	26'548	0.81%	89%
londra	13'732	0.42%	123%
other	3'018	0.09%	
total		**17.72%**	
Chicago			
tag	**number**	**% of tags**	**C. of var.**
chicago	216'969	18.59%	21%
windycity	1'559	0.13%	479%
chicagoland	1'470	0.13%	359%
chi(town)	1'469	0.13%	388%
total		**18.98%**	

Table B.1: Identified city toponyms among all tags in the bounding box of Zurich and among the 1'000 top-ranked tags in the bounding boxes of London and Chicago

City toponym tags

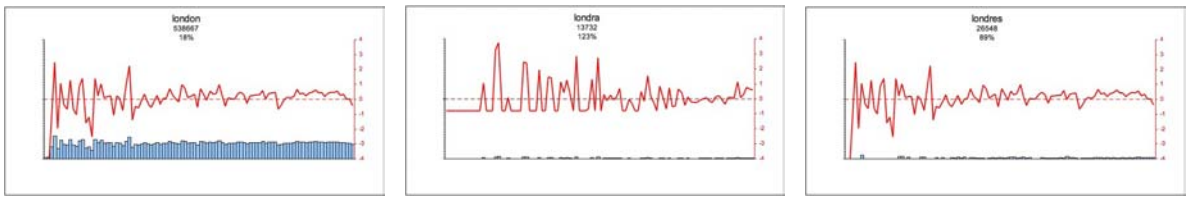


(a) zurich (b) zürich (c) zuerich



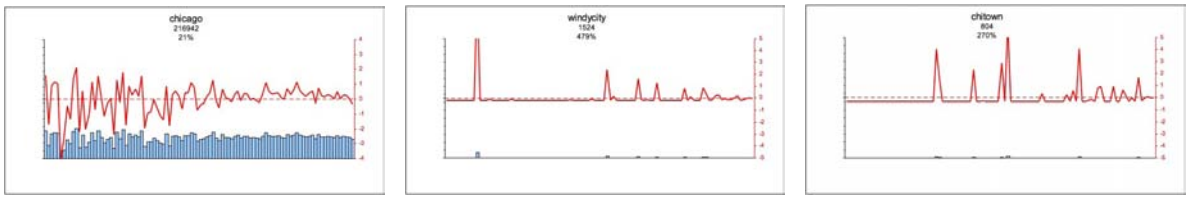
(d) zurigo (e) geo:city=zurich

Figure B.4: Tag profiles for city toponyms within bounding box of Zurich

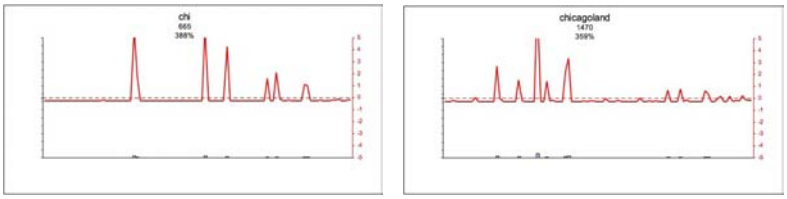


(a) london (b) londra (c) londres

Figure B.5: Tag profiles for city toponyms within bounding box of London



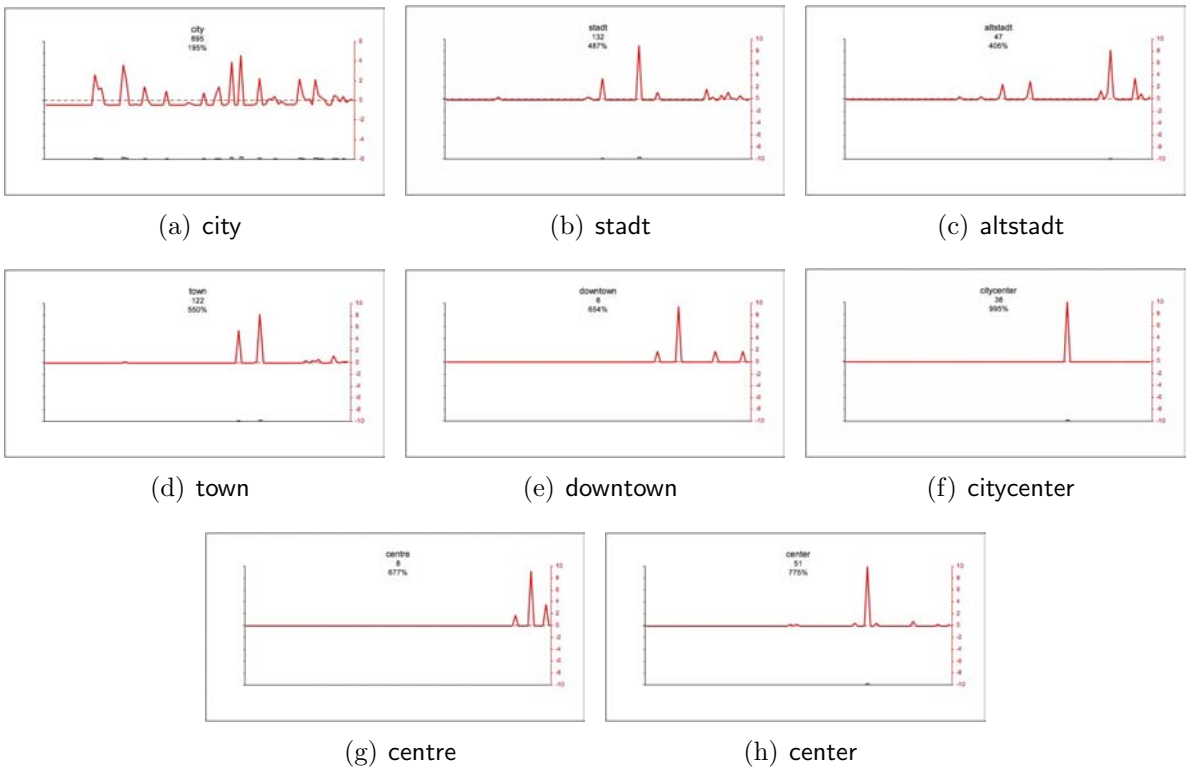
(a) chicago (b) windycity (c) chitown



(d) chi (e) chicagoland

Figure B.6: Tag profiles for city toponyms within bounding box of Chicago

Vague place tags

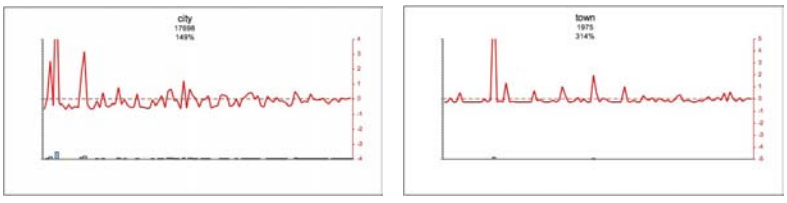


(a) city (b) stadt (c) altstadt

(d) town (e) downtown (f) citycenter

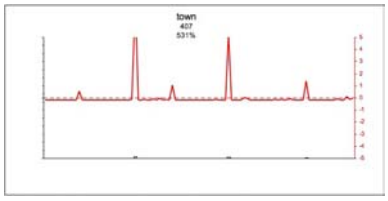
(g) centre (h) center

Figure B.7: Tag profiles for vague place tags within bounding box of Zurich

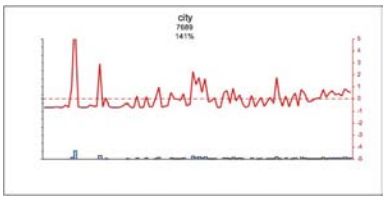


(a) city (b) town

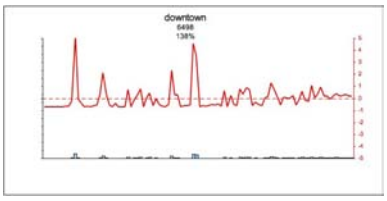
Figure B.8: Tag profiles for vague place tags within bounding box of London



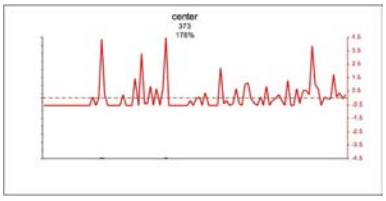
(a) town



(b) city

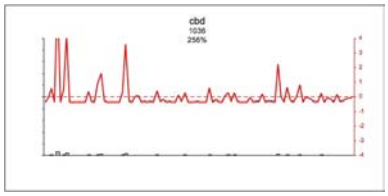


(c) downtown

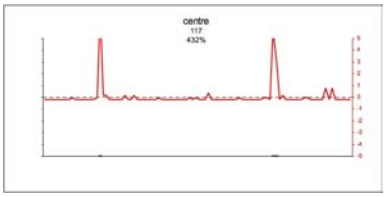


(d) center

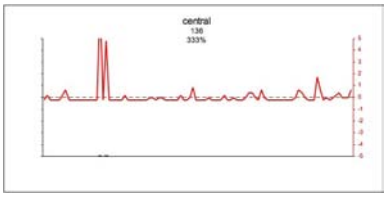
Figure B.9: Tag profiles for vague place tags within bounding box of Chicago



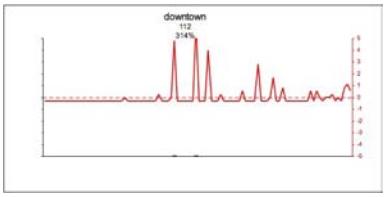
(a) cbd



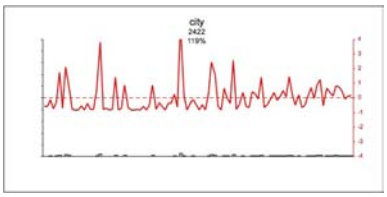
(b) centre



(c) central



(d) downtown



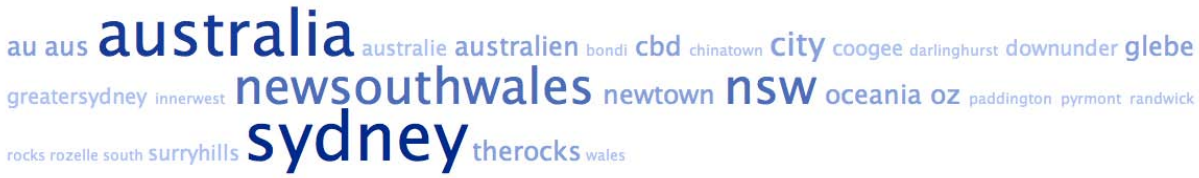
(e) city

Figure B.10: Tag profiles for vague place tags within bounding box of Sydney

Tag clouds

(a) London bbox

(b) Chicago bbox



(c) Sydney bbox

Figure B.11: Clouds of 30 most frequent specific and generic place tags occurring within the bounding boxes of different cities

B.3 Related-tag analysis

center
macro flower yellow nature world nyc pink trade white newyork flowers rockefeller newyorkcity manhattan red wtc closeup orange petals purple green rose worldtradecenter garden ny towers black naturesfinest color spring blue pollen daisy plant abigfave stamen excellence twin impressedbeauty close summer bw lily poppy

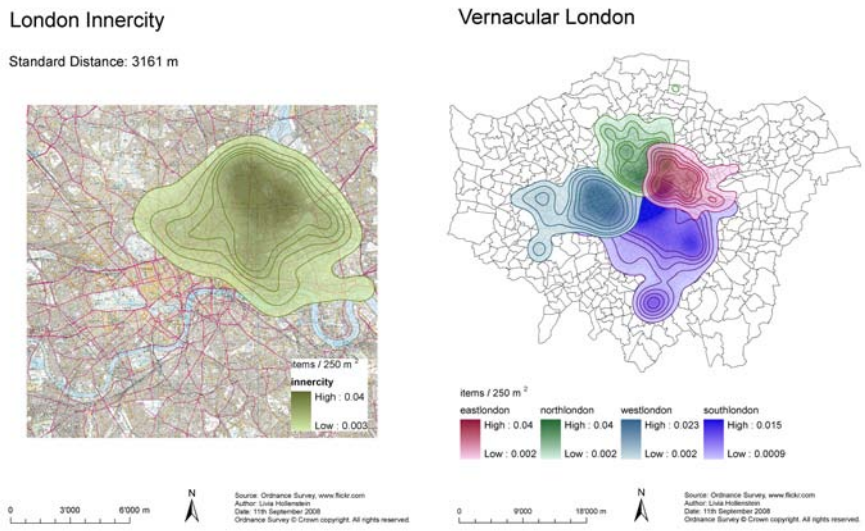
center
city macro building architecture toronto flower paris pompidou france night shopping petals uk canada england glasgow center blue ontario sky yellow white art scotland urban street pink museum mall science old orange tower purple town modern buildings london red garden canon closeup eaton naturesfinest ireland bw nature downtown island lights manchester reflection people rogers flowers river sun dublin eos water road clouds europe view bridge church rose light black sunset

center
urban street night building sky architecture buildings bw skyline people newyork downtown nyc light blue red sunset cityscape black lights white bridge art manhattan newyorkcity blackandwhite reflection clouds river longexposure dark road water cars graffiti car skyscraper cloud window green tower color windows yellow canon traffic wall landscape travel ny moon

center
city street old sky building architecture night urban buildings bw house blue travel church white people africa cape water light south europe landscape clouds bridge red black road italy river uk london england dark car art green window germany italia wall nature lights sunset downtown view cloud sea tower houses canon camden espaa spain sun square trees tree france reflection deutschland nikon blackandwhite boat summer color

B.4 Vague footprints

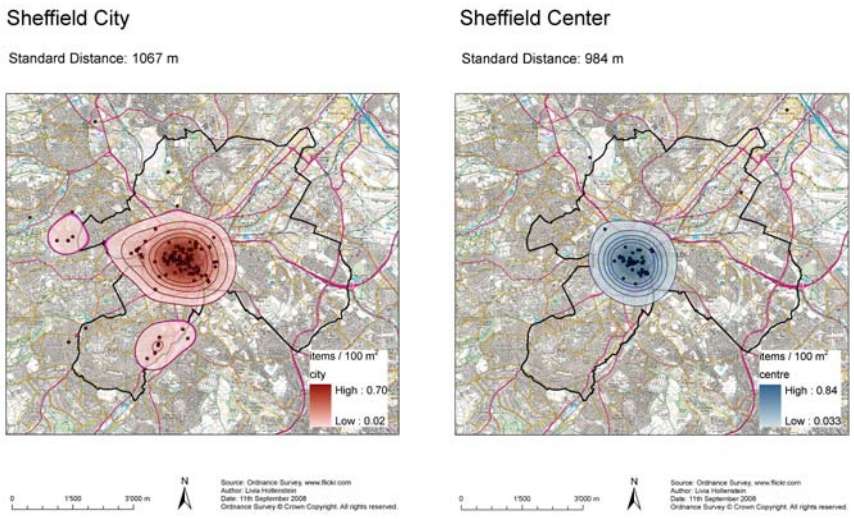
B.4.1 London



(a) (b)

Figure B.12: Vague footprints for vernacular regions of London

B.4.2 Sheffield



(a) (b)

Figure B.13: Vague footprints for place tags in Sheffield

B.5 Map data for comparison

Zurich

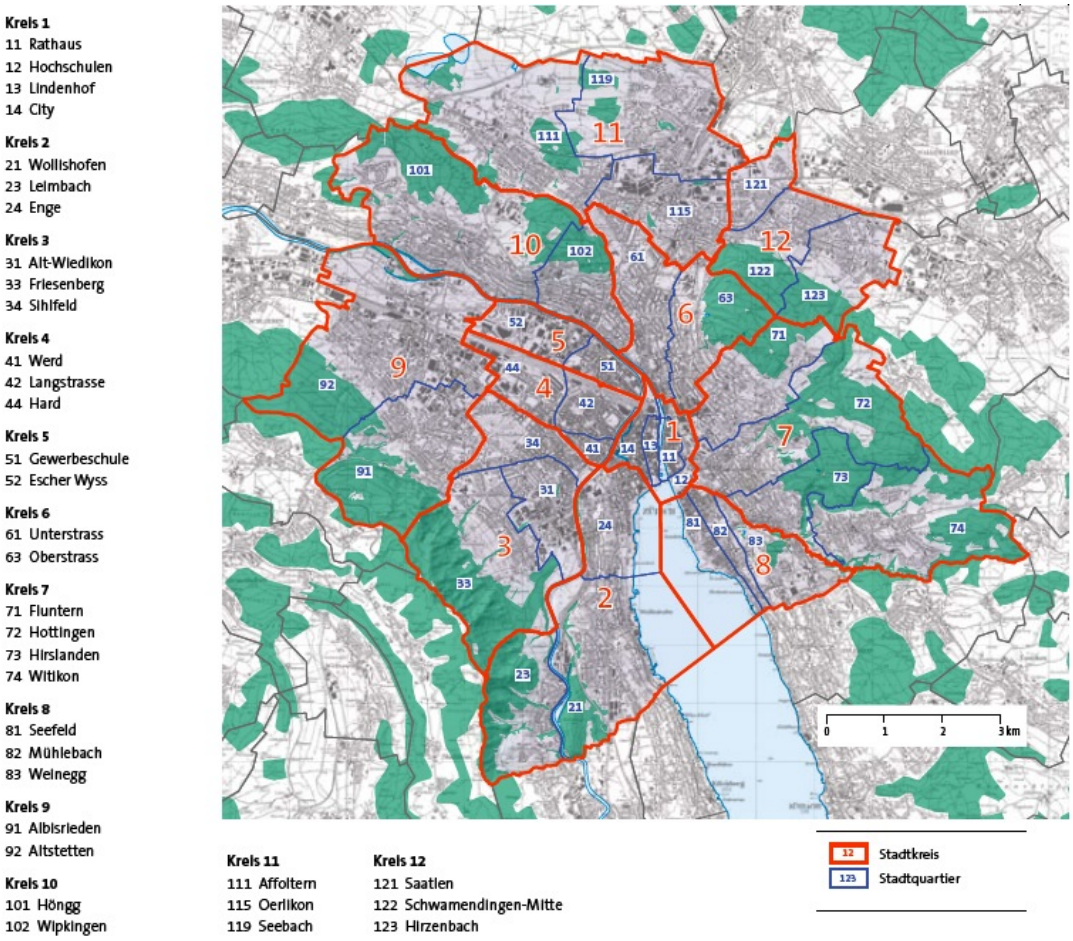
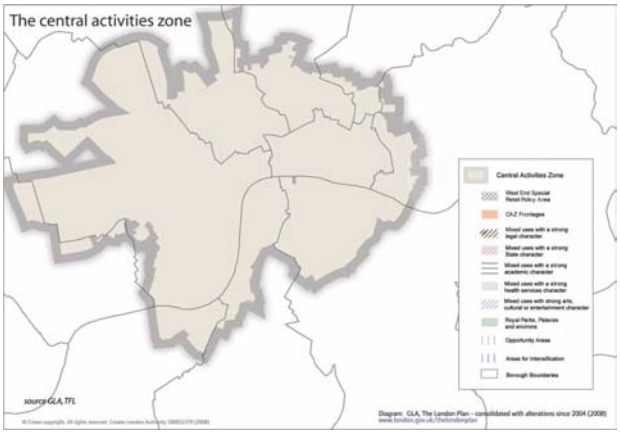


Figure B.14: Official districts ('Kreise') and neighbourhoods ('Quartiere') of Zurich (Source: Schönauer (2007: 16))

London



(a) Central activities zone

(b) Zone of congestion charge

(c) Zone 1 of the Underground

Figure B.15: Central activities zone as conceptualised by the London Plan[1] (a), the zone of congestion charge[2] (b), and from zone 1 of the London Underground[3] (c)

[1] http://www.london.gov.uk/thelondonplan/, accessed 18th October 2008
[2] http://en.wikipedia.org/wiki/Image:London_congestion_charge_zone.png, accessed 18th October 2008
[3] http://en.wikipedia.org/wiki/Image:London_Underground_Zone_1.png, accessed 18th October 2008

Sheffield

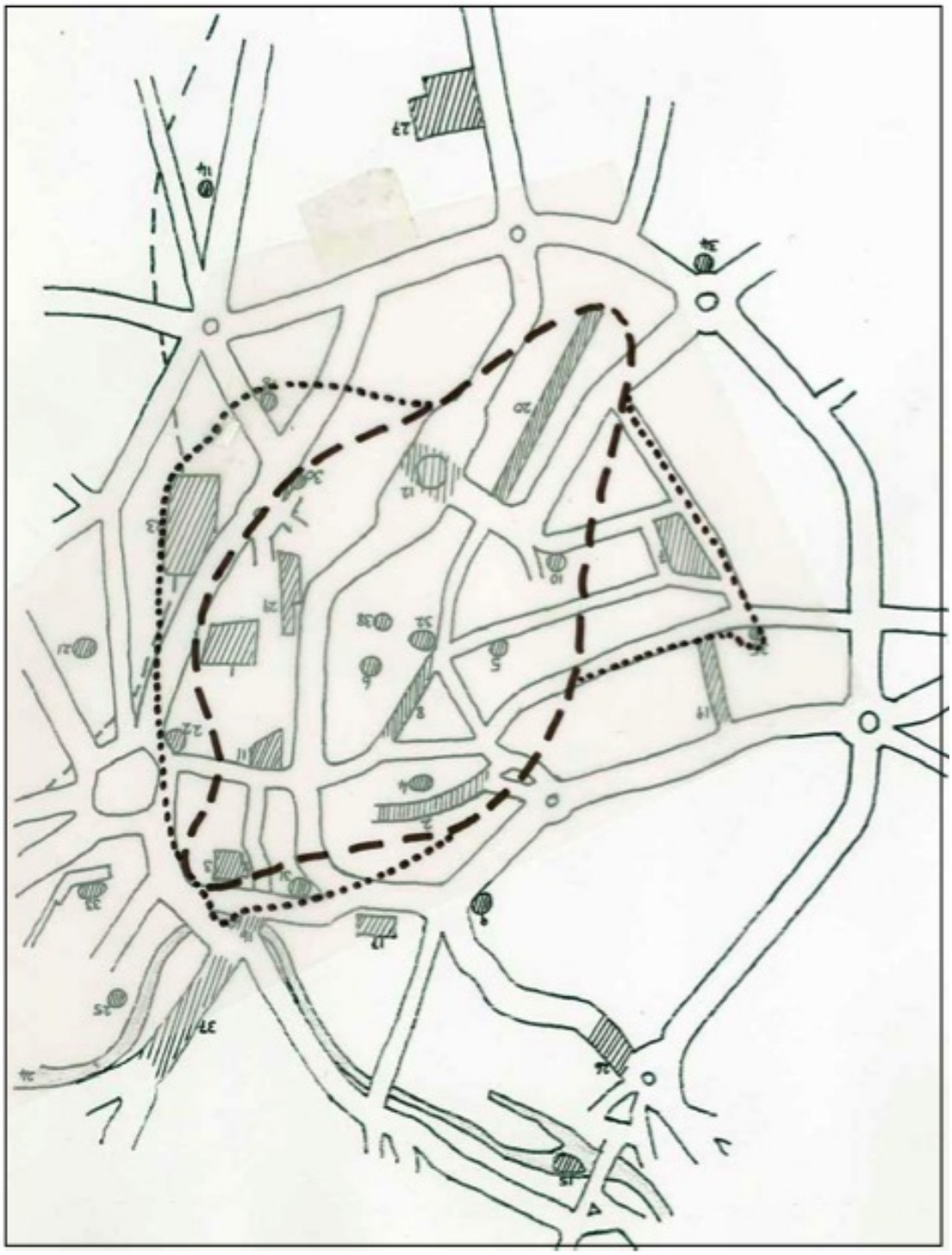


Figure B.16: City centre of Sheffield as established by Mansbridge (2005)

Chicago

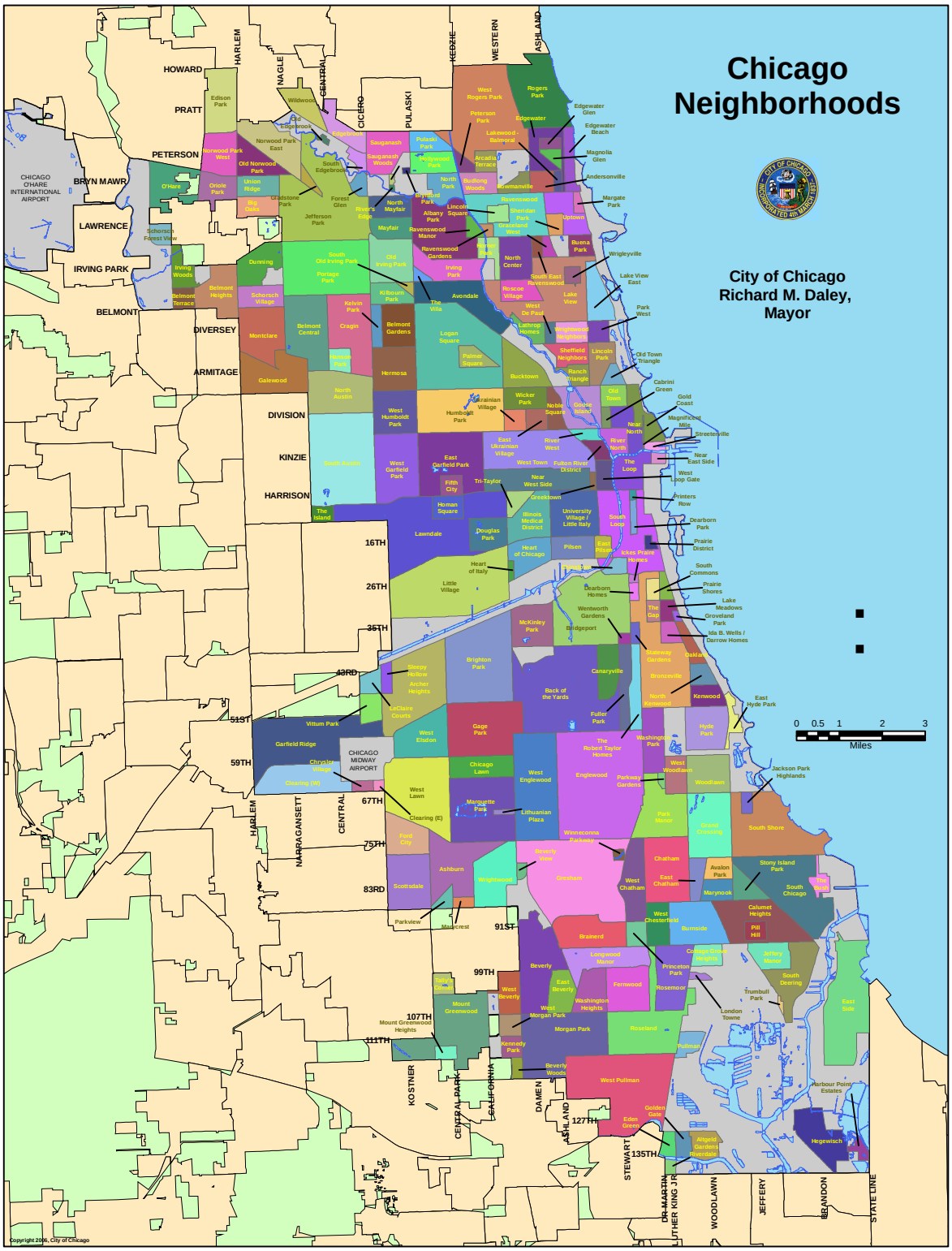


Figure B.17: Official neighbourhoods of Chicago (Source: http://egov.cityofchicago.org/webportal/COCWebPortal/COC_EDITORIAL/City_Neighborhoods_8_5x11.pdf)