

Citizens as Database: Conscious Ubiquity in Data Collection

Kai-Florian Richter and Stephan Winter

Department of Infrastructure Engineering
The University of Melbourne
{krichter,winter}@unimelb.edu.au

1 Users as Data Providers

Crowd sourcing [1], *citizens as sensors* [2], *user-generated content* [3, 4], or *volunteered geographic information* [5] all describe a relatively recent phenomenon that points to some dramatic changes in our information economy. Users of a system, who often are not trained in the matter at hand, contribute data that they collected themselves without a central authority managing or supervising the data collection process. The individual approaches vary and cover a spectrum from conscious user actions (‘volunteered’) to passive modes (‘citizens as sensors’). *Volunteered* user-generated content is often used to replace existing commercial or authoritative datasets, for example, Wikipedia as an open encyclopedia, or OpenStreetMap as an open topographic dataset of the world. Other volunteered content exploits the rapid update cycles of such mechanisms to provide improved services. For example, fixmystreet.com reports damages related to streets; TomTom and other dataset providers encourage their users to report updates of their spatial data. In some cases, the database itself is the service; for example, Flickr allows users to upload and share photos. At the *passive* end of the spectrum, data mining methods can be used to further elicit hidden information out of the data. Researchers identified, for example, landmarks defining a town from Flickr photo collections [6], and commercial services track anonymized mobile phone locations to estimate traffic flow and enable real-time route planning.

In short, user-generated content drastically reduces the costs of the data collection process and the time to next update. Since the data is user-generated and shared, and reproduction and distribution costs are close to zero [7], users expect free access. In addition, user-generated data has proven to match the quality of authoritative or commercial datasets [8]. In studying the motivation of people to contribute, Benkler demonstrates “that the diverse and complex patterns of behavior observed [...] are perfectly consistent with much of our contemporary understanding of human economic behavior” [9, p. 91], and thus, the phenomenon of user-generated content is not going away soon. The traditional economy of spatial information is fundamentally challenged.

Thinking this emergent field further into the future, some limitations become immediately obvious. User-generated spatial data consists of a combination of sensor data (e.g., coordinates) and user-added semantics (e.g., place descrip-

tions). While databases are good at collecting and interpreting sensor data, the management and use of the user-added semantics is still in its infancy.

Traditionally, semantics is inferred by data mining methods on sensor data, and the same has been done for user-added semantics. We believe that this traditional approach will fall short, since, fundamentally, adding the kind of semantics discussed in this paper is a human intelligence task. Human Intelligence Tasks (HIT) are simple, often menial tasks (e.g., identifying objects on a photo), which appear to be ideally suited to be solved automatically, but are in fact easier and cheaper to be solved through crowd-sourcing. The term was introduced by Amazon’s Mechanical Turk (<https://www.mturk.com>).

The kind of semantic information envisioned here goes well beyond the simple tasks of Mechanical Turk, so it is bound to be contributed by humans. Accordingly, the true challenge is collecting highly redundant user-added semantics—and getting enough users involved to make it highly redundant—and developing novel information inference processes for semantically rich data.

2 The Volunteered Image of the City

Crowd sourcing, user-generated content and volunteered geographic information are all ideally suited to collect *semantic information* rather than just data. In the spatial domain, locals provide not only their tracks or locations, but also their local expertise for free. For example, users map their neighbourhoods in OpenStreetMap, or annotate places in their neighbourhood in CityFlocks [10], which can then be accessed by others for local decision making. Because contributors are in the environment when they trace geometries of geographic features and describe their semantics, they are ideal candidates to provide their insights and experience for others to use: physical presence increases their credibility when reporting on the experience of spatial and social structures of environments.

People’s local experience is captured in the idea of the *image of the city* [11]. This ‘image’ is a metaphor for people’s mental conceptualisation of the environment they are living in—its spatial layout in a narrow sense and, more broadly, how they conceive the environment’s social structure. It evolves over time through interaction with the environment. It is this image that lets us find our way around, know which places to avoid at night, recommend places to buy specialist items to others, and figure out that our favourite restaurant and favourite coffee place are really only two blocks away from each other.

We all carry this ‘image of the city’ with us all the time—in our heads: the *citizen as a database*. Imagine it could be externalised! Suddenly, everybody else could benefit from our experiences. If we could store each individual’s ‘image of their city’ in electronic form, i.e., represent it in a data structure, this would open up opportunities for location based services and man-machine communication that are unheard of today. Such future services would replace asking the ‘local expert.’ In fact they would access the collective wisdom of multiple local experts at any location, which would filter our individual eccentricities. In a holistic way, these externalised images would deliver topographic data of the neighbourhood

and place annotations as in CityFlocks [10], and also provide access to local and global landmarks for navigation and gazetteers of official and vernacular place names, to name just a few examples. More generally, these services could communicate like humans [12], which would be vastly beneficial since humans are not necessarily experts when it comes to communicating with machines.

This vision of volunteering the image of the city has implications for data collection (how to capture the images), for data maintenance and integration (how to represent and map their semantics), for querying and analysis (how to exploit their semantics), and for communication (how to express semantics). Further, capture and communication require careful design of the human-computer interaction. For reasons of space in this paper, we will discuss only capturing of such images of the city in more detail. Already this limited discussion will illustrate some of the long-term research questions that come with this vision.

The problem can be phrased as: How to externalise the experientiable and experienced image of the city? And how then to put it into a (crowd-sourced) content platform for sharing with others? In contrast to general user-generated content, for spatial content users have to be or been in-situ to document their experiences, i.e., they report in real-time or from memory. Both ways, in-situ or in hindsight, have their advantages. In-situ collection may be more reliable and less distorted than from memory, but collection from memory had already a filter on selection and relevance. This interplay of capturing of both sensor and semantic information of corporeal experiences through crowd-sourcing methods is summarised as *conscious ubiquity*.

3 Conscious Ubiquity in Data Collection

Great advances in sensor technology and mobile computing allow for running powerful software while on the move. The ubiquity of mobile Internet provides means for data communication. The technologies of Web 2.0 enable users to contribute content using sophisticated (web-based) interfaces. These are the foundations for the vision of smart mobile devices capturing the image of the city.

Collecting data for this image requires a balance between ubiquity of service and dedicated human-computer interaction. People will contribute such content in large numbers and over longer periods of time only if this collection is facilitated unobtrusively, casually, or, as Weiser put it, calmly [13]. Data collection needs to be supported by automated capturing processes in order for users to accept this service as they cannot be constantly involved. Then again, externalising the image of the city requires significant human interaction; it is a human intelligence task. Adding semantic information will always be an active process, at least in part. Thus, communication between users and devices must become as natural and unobtrusive as possible. Interfaces need to disappear; contributing semantic information must become a negligible task, supported by intelligent, sensor-rich devices, such that people stay “tuned into what is happening around” [13].

At the same time, the collected data is highly sensitive, many privacy issues are involved. It is crucial that users are informed about which data gets col-

lected and how it is distributed. They need to be in absolute control of their data. Keeping this control may require breaks in the seamless collection process outlined above, in order to make users aware of potential breaches of privacy.

Just as the image of the city develops over time in our heads, it will develop and change over time in external representations. New semantic information must be integrated in these representations in a consistent and coherent manner. Often, this may make existing information obsolete. But as this information is semantic and highly individual, it may not be obvious which information to replace. Issues of semantic similarity and automatic conflict resolution need to be solved, even more so if images of multiple users are to be integrated.

In summary, technologies for volunteering the image of the city need to be smart enough to collect sensor observations, provide disappearing interfaces for collection of semantic information, report to the user on request and in critical situations, contribute the collected observations to a content platform, and smoothly integrate these observations into the platform's databases. This interplay between human intelligence, calm technology, and autonomy leads to *conscious ubiquity* in data collection, and, ultimately, to the *citizen as a database*.

References

1. Surowiecki, J.: The Wisdom of Crowds. Doubleday, New York (2004)
2. Goodchild, M.: Citizens as sensors: the world of volunteered geography. *GeoJournal* **69**(4) (2007) 211–221
3. Krumm, J., Davies, N., Narayanaswami, C.: User-generated content. *Pervasive Computing* **7**(4) (2008) 10–11
4. Haklay, M., Weber, P.: OpenStreetMap: User-generated street maps. *Pervasive Computing* **7**(4) (2008) 12–18
5. Elwood, S.: Volunteered geographic information: Key questions, concepts and methods to guide emerging research and practice. *GeoJournal* **72**(3-4) (2008) 133–135
6. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: Proceedings of the 18th International Conference on World Wide Web, New York, ACM (2009) 761–770
7. Shapiro, C., Varian, H.R.: Information Rules: A Strategic Guide to the Network Economy. Harvard Business Press (1998)
8. Haklay, M.: How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design* **37**(4) (2010) 682–703
9. Benkler, Y.: The Wealth of Networks. Yale University Press, New Haven (2006)
10. Bilandzic, M., Foth, M., De Luca, A.: Cityflocks: designing social navigation for urban mobile information systems. In: DIS '08: Proceedings of the 7th ACM Conference on Designing Interactive Systems, New York, ACM (2008) 174–183
11. Lynch, K.: The Image of the City. The MIT Press, Cambridge (1960)
12. Winter, S., Wu, Y.: Intelligent spatial communication. In Navratil, G., ed.: Research Trends in Geographic Information Science. Springer, Berlin (2009) 235–250
13. Weiser, M., Brown, J.S.: The coming age of calm technology. In Denning, P.J., Metcalfe, R.M., eds.: Beyond Calculation: The Next Fifty Years of Computing. Springer, Berlin (1997)