Smart Data Infrastructure for Urban Analytics, City Science, Planning and Design

Digital data infrastructure, and the conditions that govern its creation and maintenance, have been transformed during the 50-year lifetime of **Environment and Planning B**. My own first contributions to the journal on this topic were two editorials written half a journal lifetime ago (Longley 1995; 1998). They considered data intellectual property rights and data infrastructure provision at a time when there was frequent reference to 'data bottlenecks' and academic sector access to basic georeferencing of the built environment was far from established. The editorials were written within the purview of geodemographics, or 'the analysis of people by where they live' – with its focus on neighbourhood differentiation that can be traced to the Chicago urban ecologists of the 1920s. One of my first contributions to the journal (written with Rich Harris: Longley and Harris 1999) built on these related themes by focusing on the ways in which consumer lifestyles surveys could be used to supplement and refresh census data– important because census data provide only (a) a skeletal structure for understanding the richness and diversity of human behaviour in modern societies and (b) an infrequent (decennial) snapshot of fast-changing societies and social conditions.

25 years on, digital data support more published research than ever: in its founding year Environment and Planning B mustered 11 papers, but circa 140 in 2023. Data sources have proliferated too, including many 'Smart Data' sources arising from human interactions with smart devices. But data provenance of many of today's new sources is often not fully established, while the issues of data ownership and control that govern scientific replicability are often relegated to data access statements and do not feature prominently in academic peer review. Moreover, most Smart Data are 'found', that is created as a by-product of human interactions with no known basis to generalisation to any known population. Multiple data sources are not linked at the individual level, leading to possible ecological fallacy when analysis relies upon aggregations. 'Balkanisation of the human self' (Goodchild 2015) arises where characteristics from multiple data sources cannot be attributed to their common individual bearers. In such circumstances representation can become delusion and population behaviour becomes impossible to predict. Consequently, for example, inability to ascribe individual level human agency leads current 'digital twins' research to focus on inert built form rather than behavioural interactions (Fotheringham 2023).

Longley and Harris (1999) was an early attempt to enrich representation of human behaviour at the individual level without sacrificing generalisation to a known and pre-defined population. The lifestyles data derived from voluntary completion surveys that were circulated in newspapers, and newspaper readership was one of many ways of stratifying responses and reweighting them. Post stratification reweighting of a single data source has never been condoned in survey research practice. But what if, in the new Balkanised world of Smart Data, the coverage of **every** dataset could be ascertained, and it could be reliably assumed that every individual had been captured by at least one or other data source? On this realistic premise and the presumption that almost everyone has a geolocated residential address, it becomes possible to create an individual level baseline population, which can be ascribed different behavioural characteristics. Assumptions about missingness or abstention

can then be used to plug omissions deemed unobserved or absent. The more datasets that are triangulated together, and the more rigorous is the triangulation of these individual level representations with available aggregate statistics, the closer is the foundational digital twin to observable reality.

Realisation of this panacea is not helped by adverse data owner perceptions of disclosure control, particularly of individual level data that are defined as Personal or Sensitive Personal under General Data Protection Regulation (GDPR). Such perceptions can frustrate linkage of records from multiple sources, despite the existence of GDPR research derogations and evidence that linkage can be effected using trusted research environments (TREs). Research outputs created in such environments are managed by accredited 'safe researchers' to ensure disclosure control. However, the variable costs of output checking for individual projects poses intense resource challenges to the emergent mass research culture in urban analytics and city science. There is thus a demonstrable need for clear thinking about data infrastructure if it is to be efficient, effective and safe to use by an ever larger and more diverse user base.

Data hungry members of a mass research culture require both specialist and more generalist support. Individual researchers may happen upon a smart data source and negotiate personal access to it, but this neither guarantees scientific replicability of findings by independent researchers nor provides any obvious route to use of the data by others. Data provenance remains unchecked and further data use wastes effort in reworking bilateral data licencing agreements. No new research culture develops around new and emerging data sources.

FAIR (findable, accessible, interoperable and reusable) data access therefore needs to be negotiated through sector-wide (multilateral) data licencing of raw data and pre-processed 'research ready data' (RRD). RRD creation entails streamlined use of widely accepted procedures to prepare data for a range of research purposes – as already practiced in remote sensing, for example, where procedures applied to land surface reflectance characteristics yield content about land use for a wide range of applications. However, pre-processing of Smart Data for use across social science requires attention to coverage – **who** is and who is not included in a data source – as well as upon content – **what** the data source contributes to analysis – because the source and operation of bias in most Smart Data is not understood. This requires effective data curation and data documentation, typically achieved by triangulating Smart Data with conventional statistical sources to communicate provenance and the likely range of applications for which RRD may be applicable.

This should be undertaken using expertise in data centres familiar with statistical framework sources as well as the vagaries and uncertainties of emerging Smart Data. Accumulated experience at such centres can create **foundational Smart Data infrastructure**, the provenance of which is understood. It should pertain to people and to places and retain the human individual as the unit of data assembly wherever possible. This will both facilitate aggregation at any convenient scale in RRD products and make explicit the nature of incompleteness in representation at the individual level. Updates should be scheduled at regular (e.g. annual) intervals. Where necessary, each component dataset should be licenced

for use by the broadest possible constituency of users – subject, if necessary, to provisions to deal with commercial conflicts of interest. This underpinning infrastructure should be governed by multiple data licencing agreements (DLAs).

This infrastructure should be used to create derivative **core RRD** with content and geographic coverage capable of supporting the widest range of envisaged research projects. Prioritisation of data themes should be gauged by monitoring existing unmet user needs and the composition of current user demands. Any combination of core data might be concatenated and conflated into non-disclosive data products. Value can then be leveraged by applying widely accepted and peer-reviewed procedures. This may entail collaboration with established data providers, such as national statistical agencies, to ensure adherence to proven procedures of database creation, maintenance and usability. Updates of each RRD product should be scheduled at standard intervals, enabling both timely research and time series analysis. The resulting data may be made available to prospective users under standard terms and 'public good' conditions.

Licencing restrictions on RRD creation may mean that it will not always be possible to cross classify every data holding with all others. In such circumstances, new data sources might still be ingested into the data service secure environment for validation purposes by triangulation with other data assets. The derivative **third party RRD** can then be exported and provisioned for research applications.

A quarter of a century ago, the core UK intellectual property rights affecting urban analytics in the UK concerned ownership and control of georeferencing standards and cost recovery for new digital products such as national mapping agency or mail delivery address registers. These core infrastructural requirements have been managed and, for the academic sector at least, rendered basic digital infrastructure available under 'one to many' licencing agreements. The wider issue of how to populate this infrastructure with human characteristics and guilt environment attributes is now of concern. What is required is:

- multilateral academic sector-wide data licences, with additional provisions for cross sector research with central/local government and the business sector wherever possible
- Smart Data metadata that build upon accumulated academic and statistical agency experience of establishing data coverage and content, allied to data dialogues with providers
- Responsiveness to research user needs, in the context of increasing levels of data driven research that is interdisciplinary and cross sector with business and government
- value proposition analysis of direct linkage between core infrastructure and new sources of Smart Data
- cost – benefit analysis to foster wider academic adoption of promising new data sources.

Fotheringham, A. S. 2023. Digital twins: The current “Krays” of urban analytics? **Environment and Planning B: Urban Analytics and City Science**, *50*(4), 1020-1022. https://doi.org/10.1177/23998083231169159

Goodchild M F 2015. *Four thoughts on the future of GIS*. ArcWatch https://www.esri.com/about/newsroom/arcwatch/four-thoughts-on-the-future-of-gis/. Accessed 22 December 2023.

Longley P A 1995. Intellectual property rights and digital data Editorial **Environment and Planning B: Planning and Design**, 22: 505-8. https://doi.org/10.1068/b220505

Longley P A 1998. GIS and the development of digital urban infrastructure. **Environment and Planning B: Planning and Design**, 25: 53-6

Longley P A, Harris R J 1999 Towards a new digital data infrastructure for urban analysis and modelling. **Environment and Planning B: Planning and Design**, 26: 855-878