17 Information and Decision Making

LEARNING OBJECTIVES

After studying this chapter, you will understand:

- The role of information in decision making.
- Trade-offs, uncertainty, and risk in decision making.
- Characteristics of information and GI.
- Added value through GI linkage.
- Different types of GI.
- Open Data, Big Data, and Open Government.
- An example of a major information infrastructure.
- Where to go for more detailed information and advice.

This chapter sets out the ways in which geographic information (GI) can make it possible to use the techniques described in earlier chapters in order to make sound decisions. GI is considered both as a good and as a value-adding service. We review the economic and other characteristics of GI and consider how different data types and business drivers have implications for GI system use. We highlight the numerous trade-offs and uncertainties inherent in using GI for decision making and the ways in which value is added through data linkage. We describe the implications of Open Data concepts and practice for users and developers of GI technology functionality in many countries. Finally, we use a military example to illustrate how all of these can be brought together in an information infrastructure created to aid decision making.

The discussion complements the "hard science" perspective of previous chapters by explaining the relevance of economics, human behavior, public policy, and sometimes politics in GI management. The objective of efficient and effective management is to arrive at and implement decisions without losing public trust, incurring excessive costs, or being challenged successfully by lawyers.

17.1 Why We Need Information

Most readers of this book live in capitalist societies, with all their advantages and disadvantages (see Chapter 19). In such societies, manufacturing has diminished as a source of employment, and information-based services have become dominant. According to the UN Economic Commission for Europe, in 2011 the median proportions of the workforce employed in some 20 developed countries was approximately 3% in agriculture, 20% in industry, and 75% in services. Even in countries where agriculture and manufacturing industry are still strategically important employers, there is movement up the value chain associated with increasing employment in information-based services. Much of that information is geographic in nature. It tells us where things are happening, where there is high attraction to live

and work, where a critical mass of specialists operates or natural resources exist, which is a generally safe area, how to get from point A to point B, and so on. GI systems as services (Section 1.5.2.3) already have great value and are growing rapidly. Figure 17.1 summarizes one estimate of these areas carried out for Google by a respected economic consultancy. This estimates that the global geospatial services sector generates $150 to $270 billion annually. The Boston Consulting Group estimated that the U.S. geospatial

Figure 17.1 The economic impact of geo-services. (Source: Oxera, 2013)

Technical Box 17.1

Some uses of geographic information

These include

- Describing the current status of some phenomenon (e.g., the geographic distribution of population), the historical changes in that distribution, or projections of what will happen to it in the future. The identification of geographic patterns can suggest possible explanations.
- Seeking reasons why the distribution is as it is:
  - By establishing correlations between it and other variables.
  - By seeking to isolate causality.
- Acting on the results by making decisions or facilitating decision making by others, based on an understanding of the processes that led to the distribution or that will create a different pattern in future.

industry generated some $73bn in 2011 and was composed of at least 500,000 jobs. Box 17.1 sets out some of the specific objectives we have in using geographic information.

Management without relevant information is like driving in the dark without headlights.

But even if the "hidden hand" of capitalism shapes how societies evolve and prosper, most of us also live in some form of a managed economy—whether this is in China, Russia, France, or many other countries. Managers exist to make decisions and implement them successfully. To do this requires analytical tools such as multicriteria decision-making techniques (MCDM: Section 15.4), but these tools need to be fed with information. Management without relevant information is like driving in the dark without headlights. Increasingly, there is a demand to ensure that decision making in managed economies relies on demonstrably good evidence. In some countries there is also pressure to ensure that the evidence is available for widespread scrutiny in order to hold governments to account (Sections 1.5.3 and 17.4). Every organization—businesses, governments, voluntary organizations, and even universities—has managers, and everyone at some stage is going to be a manager in some function. Given that much of the information we use is geographic, it follows that GI and GI systems are central to many management functions and much decision making.

Geography shapes decisions and their consequences.

17.1.1 Trade-Offs, Uncertainty, and Risk

Good decision making can be difficult and entails taking account of many environmental or contextual factors as well as simpler operational ones (Chapter 18). All decisions involve trade-offs—for example, the relatively simple personal decision of whether to purchase a more expensive house if that reduces the commute to work. More generally, a trade-off almost always exists between the benefits and disbenefits of any decision and its subsequent implementation. Sometimes these decisions bring near-universal benefits (like the decision to build GPS; see Box 17.7). But in many cases we are operating in a zero-sum game or something close to it. Benefits accrue to one group and disbenefits fall on another (at least in the short term). Sometimes these benefits or disbenefits provide advantages only to a small number of people such as the owners selling a GI services enterprise. Sometimes the benefits accrue to the whole of society, though at some cost to something else that society holds dear. In some cases the benefits take a long period to appear, whereas costs are experienced immediately. And finally, the balance between private gain and public benefit that is acceptable differs in different societies; there is a geography of trust, privacy, openness, tolerance, and commercial exploitation founded on different cultures and laws.

Everything is interconnected; usually, some gain and some lose when a decision is made.

In the real world, rarely is geography the *only* important issue with which managers have to grapple. And rarely do managers have the luxury of being solely in charge of anything or have all the information needed to make truly excellent decisions. The reality is about operating in a situation where knowledge of the options and the likely outcomes of different decisions is incomplete. Uncertainty of many types is inevitable, and hence risk exists and needs to be mitigated (see Chapters 5, 16, and 19).

17.1.2 Organizational Drivers

Organizations and individuals operating in different sectors have some shared and some different

drivers for their actions. Table 17.1 summarizes these in a simplified way for the business and government sectors.

Over the past 20 years we have seen some breakdown of previously discrete sectors. There is some convergence between the activities and operations of commerce and industry, government, the not-for-profit sector, and academia. Some governments have outsourced some of their functions (e.g., the operations of utilities). Increasingly,

Table 17.1 Some business drivers and typical responses. Note that some of the responses are common to different drivers, and some drivers are common to different types of organization.

Sector	Selected Business Drivers	Possible Response	GI-Related Example
Private	Create bottom-line profit and return part of it to shareholders. Build tangible and intangible assets of firm. Build brand awareness.	Get first-mover advantage; create or buy best possible products, hire best (and most aggressive?) staff; take over competitors or promising start-ups to obtain new assets; invest as much as needed in good time; ensure effective marketing and "awareness raising" by any means possible; reduce internal cost base.	Purchase and exploitation of MapGuide software by Autodesk and subsequent development—one of the earliest Web mapping tools. Engagement of Esri in collaboration with educational sector since about 1980, leading to 80%+ penetration of that market and most students then becoming Esri software-literate. Purchase of GDT by Tele-Atlas to obtain comprehensive, consolidated U.S. database and to remove major competitor.
	Control risk.	Set up risk management procedures, arrange partnerships of different skills with other firms, establish secret cartel with competitors (illegal)/gain de facto monopoly. Establish tracking of technology and of the legal and political environment. Avoid damage to the organization's "brand image"—a key business asset.	Typically, GI service firms will partner with other information and communication technology organizations (often as the junior partner) to build, install, or operate major information technology (IT) systems. Many GI system software suppliers have partners who develop core software, build value-added software to sit "on top," act as system integrators, resellers, or consultants. Data creation and service companies often establish partnerships with like bodies in different countries to create pan-national seamless coverage. Know what is coming via network of industry, government, and academic contacts.
	Get more from existing assets.	"Sweat assets," e.g., find new markets, which can be met from existing data resources, reorganized if necessary.	Target marketing: use data on existing customers to identify like-minded consumers and then target them using geodemographic information systems.
	Create new business.	Anticipate future trends and developments and secure them.	Go to relevant conferences and monitor developments, for example, via competitors' staff advertisements. Buy start-ups with good ideas (e.g., Esri and the City Engine technology). Network with others in GI industry and adjacent ones and in academia to anticipate new opportunities.

Table 17.1 (*continued*)

Sector	Selected Business Drivers	Possible Response	GI-Related Example
Government	Seek to meet the policies and promises of elected representatives or justify actions to politicians to get funding from taxes.	Identify why and to what extent proposed actions will have an impact on policy priorities of government and lobby as necessary for tax appropriations. Obtain political champions for proposed actions: ensure they become heroes if these succeed.	Attempts to create national GI-related strategies and capacities, e.g., via NSDIs (see Section 18.6.1). Force the pace of progress on interoperability, e.g., to meet the needs of homeland security, minimize environmental hazards such as flooding. Bring data together from different government departments to demonstrate how funding is distributed geographically in relation to need or agreed political imperative (e.g., additional UK funding per capita in Northern Ireland).
	Protect citizens from threats, e.g., war, crime.	Ensure equipment, military and other skills and manpower, and information infrastructure is adequate to warn off aggressors or triumph in conflict.	Obtain surveillance capabilities such as UAVs (Section 17.5.1), build integrated information infrastructure using geographic locators to link information derived from SIGINT, HUMINT, OPENINT, etc. (Section 17.5.2). Ensure that human capital is adequate to keep up with fast-moving developments in GI-related areas.
	Provide good value for money (VfM) to taxpayers.	Demonstrate effectiveness (meeting specification, on time, and within budget) and efficiency (via benchmarking against other organizations).	Constant reviews of VfM of British, Norwegian, and Swedish government bodies (including British Ordnance Survey, Meteorological Office) over 15+ years, including comparison with private-sector providers of services. National performance reviews of government (including GI system users) in the United States and the impact of e-government and other initiatives.
	Respond to citizens' needs for information or for enhanced services via the Web.	Identify these; set up/encourage delivery infrastructure; set laws that make some information availability mandatory.	Setting up the U.S. National Geospatial Clearing House (1992), Geospatial One-Stop (2001), Geospatial Line of Business (2006), Geospatial Platform (2010), and Geospatial Shared Services (2013) + equivalent developments elsewhere (Chapter 18). Seek to create international level playing field for information trading, for example, through European Union INSPIRE initiative (see Section 18.6.2). Hold competitions to obtain new ideas from innovators (typically over half the entries use GI).
	Control risk.	Avoid Congressional/Parliamentary exposure for projects going wrong and media pillorying.	Get buy-in from superiors at every stage. Adhere to risk-management strategies and processes. Do numerous pilot studies in several areas.
	Act equitably and with propriety at all times.	Ensure that all citizens and organizations (clients, customers, suppliers) are treated identically and that government processes are transparent, publicized, and followed strictly.	Treat all (including freedom of information) requests for information equally. Put all suitable material on Web but also ensure material is available in other forms for citizens without access to the Internet.

everyone is concerned with explicit goals (many of which share similar traits) and with knowledge management. Moreover, in the GI world at least, we also see significant overlap in functions. For instance, both government and the private sector are important producers of geographic information. The not-for-profit sector increasingly acts as an agent for or supplement to government and often operates in a very business-like fashion. The result is that previous distinctions are becoming blurred, and movement of staff from one sector to another is becoming more common.

In practice then, virtually every organization now has to listen and respond to customers, clients, patients, or other stakeholders in their area of operations and find new ones. Every organization also has to pay attention to citizens whose power sometimes can be mobilized successfully against even the largest corporations: In the Internet era, it is easier than ever for citizens to provide feedback, such as volunteered geographic information (VGI; Sections 1.5.6 and 17.3.1.3). Every organization has to plan strategically and deliver more for less input, meeting (sometimes public) targets (e.g., profitability or service quality). Everyone is expected to be innovative and to deliver successful new products or services much more frequently, cheaply, and rapidly than in the past. All managers have to act and be seen to be acting within the law, regulatory frameworks, and some conventions. Finally, everyone has to be concerned with risk minimization, knowledge management, and protection of the organization's reputation and assets. Failure to do so can have disastrous organizational and individual consequences.

The emphasis placed on these objectives, however, differs significantly between organizations in different sectors. This translates into the parameters defining the decision-making space within which an individual operates (see Table 17.1). Thus commonality between the different sectors should not be exaggerated. For instance, mergers and acquisitions are more common in the commercial than the government sector. One example of this is the purchase of Navteq, the car guidance data supplier, by Nokia, the Finnish mobile telecommunications business, in July 2008 for $8.1 billion. Another is the $1bn purchase by Google in June 2013 of Waze, an Israeli creator of a traffic and navigation app for smartphones. Yet another is the merger of GeoEye and DigitalGlobe, two major providers of fine-resolution satellite imagery, in January 2013; this was driven by cutbacks in the U.S. federal government's budgets. But, irrespective of such acquisitions being primarily a feature of the commercial world, it is still realistic to regard all good organizations as operating in a business-like fashion.

For these reasons, we use *business* as a single term to describe the corporate activities in all four sectors identified earlier—commerce and industry, government, not-for-profit, and academia. Accordingly, we think of the GI world as being driven by organizational and individual objectives, using scientific understanding and raw material (data, information, evidence, knowledge, or wisdom; see Section 1.2), tools (GI system software and hardware), and human capital (skills, insight, attitudes, and experience) to achieve them.

17.2 Information as Infrastructure

The importance of maintaining a sound physical infrastructure—roads, railways, utilities, and so on—is well recognized. Thus the interstate highways of the United States transformed commerce, employment, and recreational travel from the 1950s on, and China's sustained growth since the 1990s has been driven by investment in physical infrastructure. In times of recession, many governments across the world see suitable "shovel-ready" infrastructure projects as a means of stimulating the economy and generating long-term efficiencies. Despite this, relatively few governments have seen the parallel between physical and information (content) infrastructures. We believe this to be a close (but not exact) parallel.

Yet a coherent interoperable national information infrastructure (NII)—of which GI is a key part—brings substantial economic benefits and competitive advantage. Such NII must utilize strategically important datasets to meet current and near-term future needs in government and commerce, consistent with relevant policies, procedures, standards, directories, metadata, tools, user guidance, and skill sets. Increasingly it will include information derived from the entire public sector—not just central or federal government—and the private sector.

The case for ensuring the existence, coherence, and quality of a national information infrastructure is made more, rather than less, important by the data deluge (see Table 1.2). Changing technology has enabled us to collect ever more data. It was claimed by reputable scientists in 2012 that 90% of all information ever created had been collected in the previous two years. Not all of this is of high quality. Some is of unknown provenance and has little metadata (Section 10.2). Examples of large-volume GI include point clouds from LiDAR, satellite imagery, and cell phone and credit card records geotagged by use of the Global Positioning System (GPS; Section 4.9 and Box 17.7).

Few of us live in some socialist planned economy where the state decides everything. The private sector is a crucial player in creating information and exploiting it. Nevertheless the state has a role to ensure that

certain key information sets exist—core reference data—and that need to be of good quality and widely (arguably freely) accessible. These act as frameworks to which other data are attached (see Section 17.2.1.4).

There is a precedent for such action: The GI community and many national governments pioneered the concept of a certain subset of NII through the development of what have been termed national spatial data infrastructures (NSDI). These are discussed in more detail in Section 18.6.

But the information community, of which the GI community is one part, is not just a random collection of sets of information and tools to handle them. Because there are many interactions between what is available and other factors, we are dealing with an ecosystem. It has biotic components (humans and their skills) and abiotic ones (data, interdependencies, network of communications, institutional arrangements). It is strongly influenced by internal and external factors such as leadership, access to data (analogous to nutrients), and financing. And it is growing in size and mutating in other characteristics. Like other ecosystems, its future is difficult to predict in any detail even though it is reproducing rapidly at present.

17.2.1 Information for Management

Information, as seen from a management perspective, has a number of unusual characteristics as a commodity. In particular, it does not wear out through use, though it may well diminish in value as time passes. On occasions its value may rise again somewhat, where it is used for historic comparisons. The U.S. government decision in 2008 to make the entire Landsat archive freely available created an opportunity for researchers across the world to analyze temporal changes in the environment over a forty-year period (see, for example, Section 19.6.7).

Information does not wear out but may become outdated.

17.2.1.1 Information as a Public Good

Information is in general a public good. A pure public good has very specific characteristics:

- Even though the initial cost of collecting, quality assuring, and documenting information may be very high, the marginal cost of providing an additional digital unit is close to zero. Thus, in effect, copying a small amount of GI adds nothing to the total cost of production (see Section 1.2); however, where large datasets and high response rates are involved, the costs (and rewards) of computing, storage, and power for data analysis, dissemination, and exploitation may still be significant. One example is high-frequency trading in the foreign exchange markets where close geographic proximity of the source of release of market-sensitive data (e.g., government statistical institutes) and the location of analysis tools can provide an opportunity for algorithms to trade many millions of dollars in very short time windows (milliseconds or less). Such activity would ensure that there is not a level playing field in a market where over $4 trillion is traded every day.

The marginal cost of providing an additional digital copy is close to zero.

- Use by one individual does not reduce availability to others (termed nonrivalry). This characteristic is summarized in the famous Thomas Jefferson quotation: "He who receives an idea from me, receives instruction himself without lessening mine; as he who lights his taper at mine, receives light without darkening me."

The use of information by one individual does not reduce availability to others.

- Individuals cannot be excluded from using the good or service (termed nonexcludability). Ways of achieving such a ban include designating information as restricted (e.g., the habitats of endangered species) or through pricing mechanisms.

In practice, information is an optional public good, in that—unlike national defense—it is possible to opt to take it or not; not everyone chooses to use freely available U.S. Geological Survey data, for example! To be pedantic, it may also be best to define information as a quasi-public good because it may be nonrival (see second bullet point listed earlier), but its consumption can in certain circumstances be excluded and controlled. The business cases and vast investments of a number of major commercial GI purveyors such as GeoEye/DigitalGlobe are based on this proposition. If everything they produced could be copied for free and redistributed at will by anyone, their business would be at risk (but different business models are used by other commercial players such as Google, as discussed in Section 18.2).

Thus the monetary value of information may depend on restricting its availability, whereas its social value may be enhanced by precisely the opposite approach—another trade-off. To complicate matters, a particular set of information is also often an "experience good" that consumers find hard to value unless they have used it before.

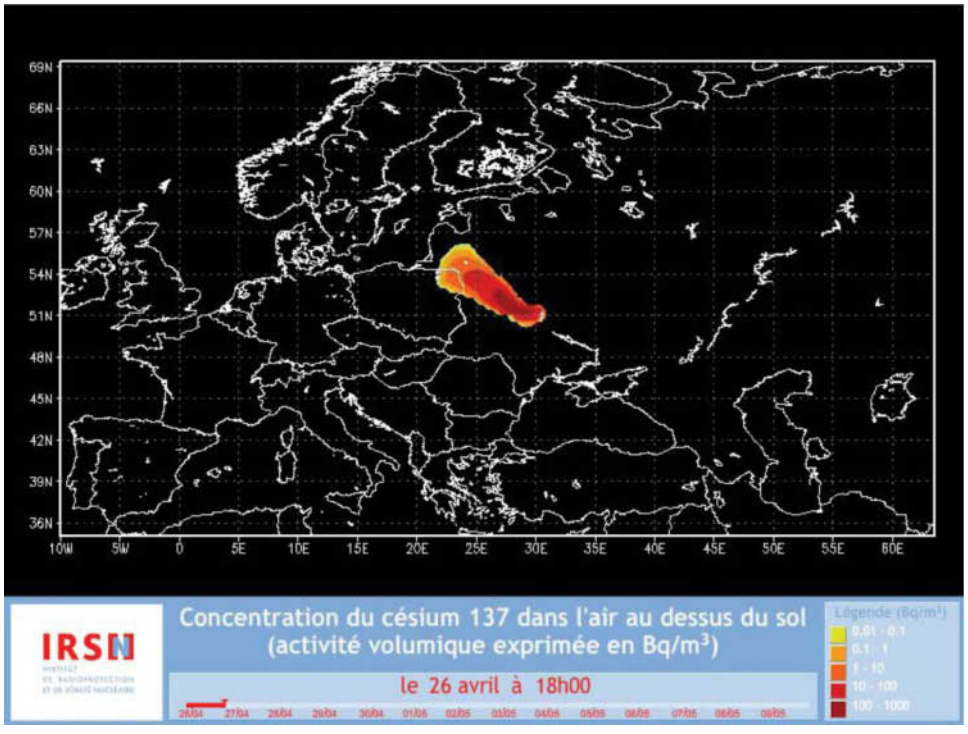
17.2.1.2 Externalities

An externality is a cost or benefit resulting from an activity or transaction that affects an individual or

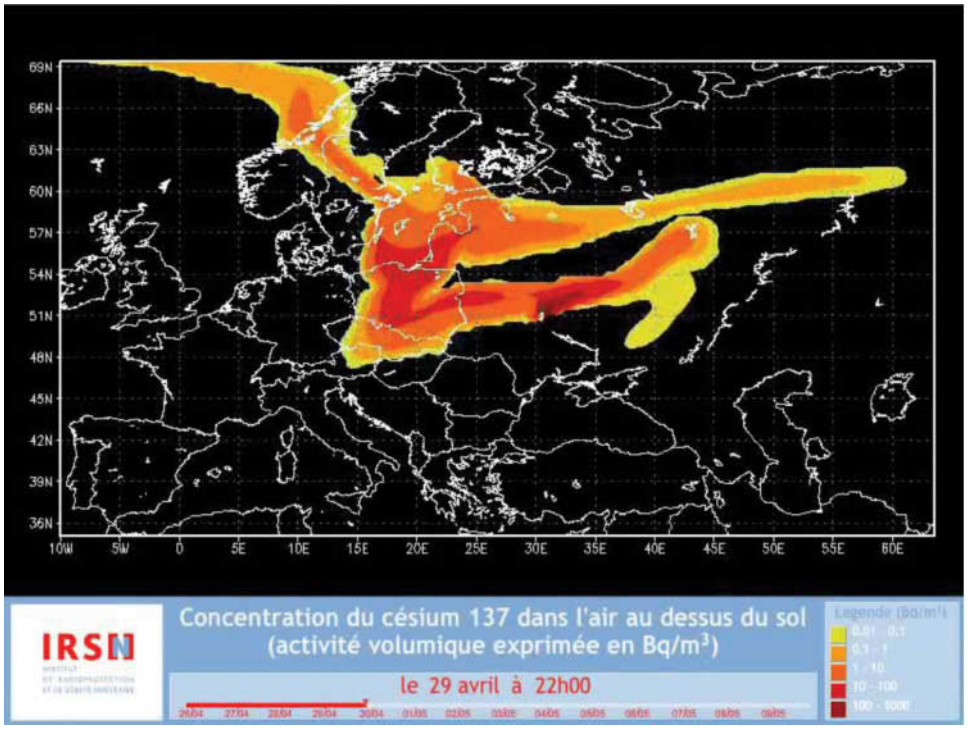
community without their direct involvement. A pure public good is a special form of externality. Negative externalities arise where production or consumption of a good (in this case, information) by one agent imposes unavoidable costs on other producers or consumers. Pollution of water or air and pollution by ambient noise are classic examples of externalities arising from external costs—that is, disbenefits. Figure 17.2 is of four video-clips showing the dispersion of atmospheric release of caesium-137 in

Figure 17.2 A to D The progress of the radioactive plume after the Chernobyl nuclear accident—a negative (geographic) externality. (Source and copyright: Institut de Radioprotection et Sûreté Nucléaire (IRSN), France) (*continued*)

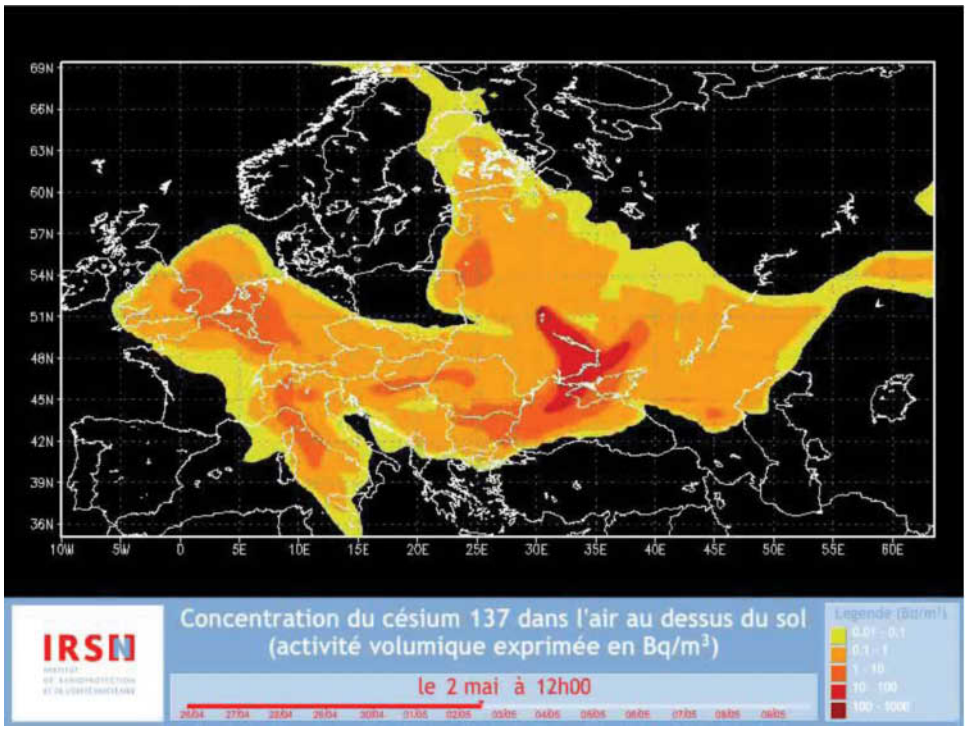
(A)



(B)



(C)



(D)

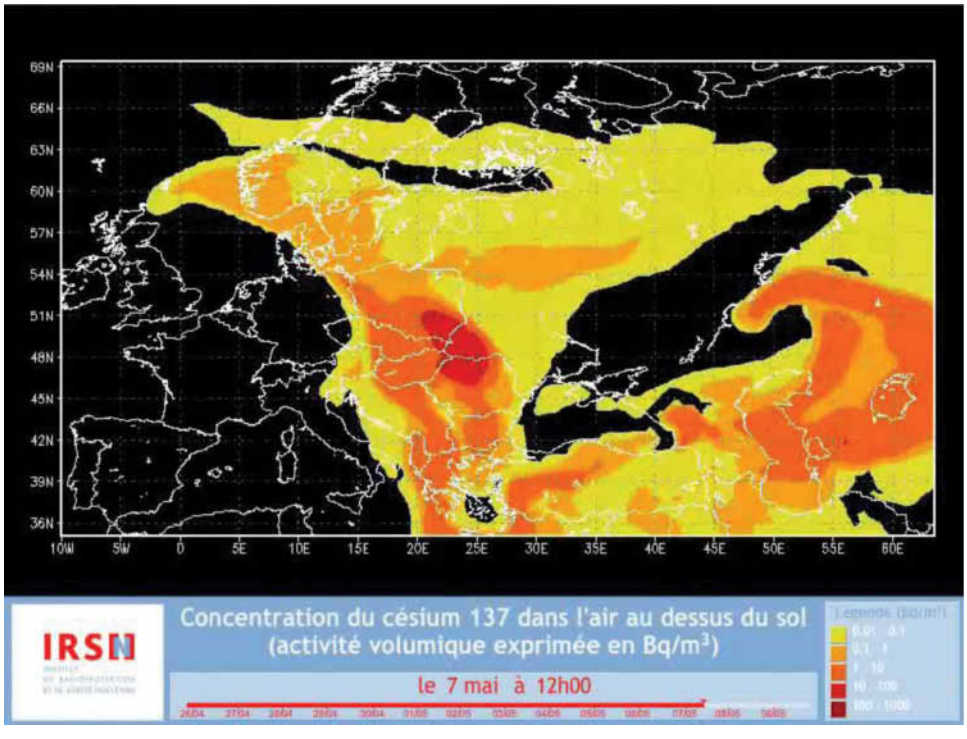


Figure 17.2 (*continued*)

becquerels/m^3 from Chernobyl. The model calculated the distribution of air contamination at ground level on a European scale at 15-minute intervals, from 26 April to 10 May 1986. There was good agreement between calculations and ground measurement data.

Classic examples of positive externalities (benefits) include refuse collection, education, and public health. Although individuals who benefit from positive externalities without paying are considered to be free-riders, it may be in the interests of society to

encourage free-riders to consume goods that also generate substantial external benefits for others.

There are different types of externalities in the information world:

- Ensuring consistency in the collection of information creates *producer externalities* by reducing the costs of creating and using data. This in turn broadens the range of potential applications.
- Providing users access to the same information produces *network externalities*. It is, for example, desirable that all the emergency services use the same geographic framework data.
- Promoting the efficiency of decision making generates *consumer externalities*. For example, access to consistent information allows pressure groups to be more effective in influencing government policy or monitoring activities in regard to pollution.

17.2.1.3 Price Elasticity and Commoditization

If information is not free to the user, price elasticity becomes an important characteristic. The demand for a good is said to be *inelastic* when changes in price have a relatively small effect on the quantity sought. Hence for some users it is critical that they must have the most up-to-date or highly detailed GI, such as emergency services. The demand for a good is said to be *elastic* when changes in price have a relatively large effect on the quantity sought. Individuals might buy many different hiking maps or datasets if the unit price declined the more they bought. Substitution effects come into play if the price is too high, when users will seek alternative sources of information.

A *commodity* is a good—such as coarse-resolution GI—that is inexpensive and extremely widely used. This has impacts on the business model of the seller (see Section 18.2). The user's choice of what to purchase is then influenced by perceptions of reliability and quality, the nature of the marketing, and convenience.

17.2.1.4 Distinctive Characteristics of GI

GI has the following characteristics, which differ in degree from other data:

- Some GI acts as frameworks to which other GI are fitted; many call these core reference data. The most basic frameworks are geodetic data and coordinate systems (Section 4.8). For many organizations and users, however, the everyday framework consists of geocoded post- or zip codes, official topographic maps and imagery, plus other GI like the distribution of population derived from censuses, etc. All such data have been sampled (Section 2.4), generalized (Section 3.8), and projected (Section 4.8) in ways that ensure that they may differ from other apparently similar GI. Where a high-quality, maintained, and widely available data framework exists, there are hugely positive network externalities (see Section 17.2.1.2) in fitting other data to it. The more people do this, the better datasets will fit each other. There is, however, a complication. In the past most of these frameworks were provided by governments. The advent of such products as Google Maps and Google Earth has changed the situation; their global availability and free noncommercial use have made them a de facto framework used by many organizations on which to "pin" their own data (see Section 4.12).

  The more people who use a framework dataset, the better datasets will fit each other and the more one source of uncertainty is reduced.

- Linking multiple datasets provides added value, at almost no cost. The number of possible overlays of two datasets rises very rapidly as the number of input datasets rises (Figure 17.3). Assuming order is not important and repetition is not allowed, one combination exists of two variables, and 1,048,555 exist with 20 variables. Many of the latter are of course likely to be of little value (i.e., they include all groups of 2, 3, 4 . . . 20 variables). Such data linkage necessitates a common linkage key. In many cases the location of the data objects provides that key. Thus many more applications can be tackled and new products and services created using linked GI than when they are held separately as multiple files.
- It is particularly difficult to quantify the quality of some types of GI, for example, area classification data like soil type. This has ramifications when combining data by overlay (Section 5.4.4) and in modeling. Quite often we have to use proxy measures of quality, such as the established reputation of the data creator in deciding which data to use. As in almost all GI system operations, we are wise to vary the inputs systematically in sensitivity analyses to see whether the results are stable (see Section 15.5).
- There is huge geographic variation in the need for and use of GI. Thus GI for urban areas is much more heavily exploited than that for rural areas—in the main because more than half of the world's

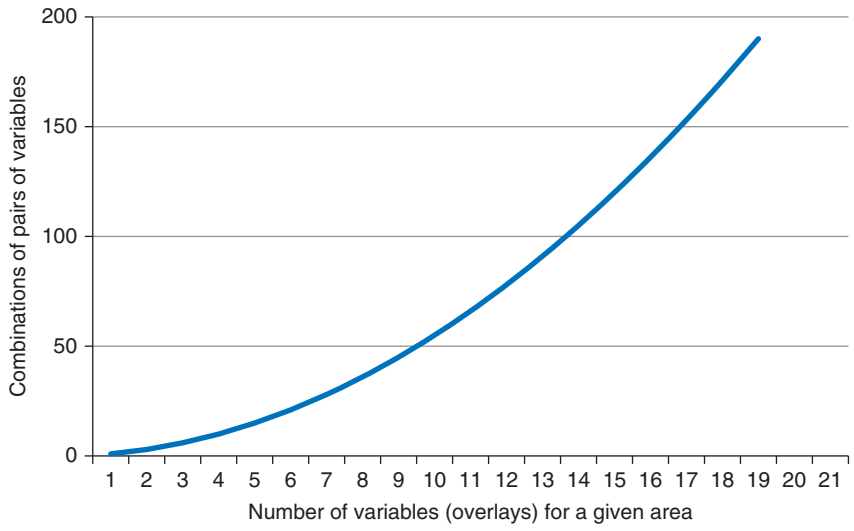


Figure 17.3 The numbers of pairs of variables that can be selected from 2 to 20 variables (= overlays) for the same area.

population now lives in cities but the urban extent of cities is probably less than 2.7% of the land area (excluding Antarctica). Nevertheless, even where there is no commercial case for provision of data in some rural areas, there is sometimes a strong social case. This may therefore require government support. Figure 17.4 makes this point: when Pan American flight 103 was brought down over Scotland by a bomb in 1988, wreckage was scattered over a 40-km-wide area. The emergency services immediately needed detailed and up-to-date maps to ensure they missed no areas in their hunt for bodies and wreckage. Such rural mapping is difficult to cost-justify because of the small number of users in normal circumstances. Here it provided a form of insurance—vital in that case.

Added value—almost for free—can be created by data linkage using location as the linkage key.

17.3 Different Forms of GI

The characteristics of available information and GI in particular shape the opportunities and pitfalls that GI system users face. We have already described some classifications of GI (see Sections 3.4 and 3.5). But all classifications are approximations to reality and are best tailored to particular purposes: none is universally useful for all purposes. Thus far we have tended to think of GI as highly structured data, typically

Figure 17.4 (A) The wreckage of Pan-American flight 103 brought down over Lockerbie, Scotland in 1988 by a bomb. (*continued*)

(A)

© Christopher Pillitz/Getty Images

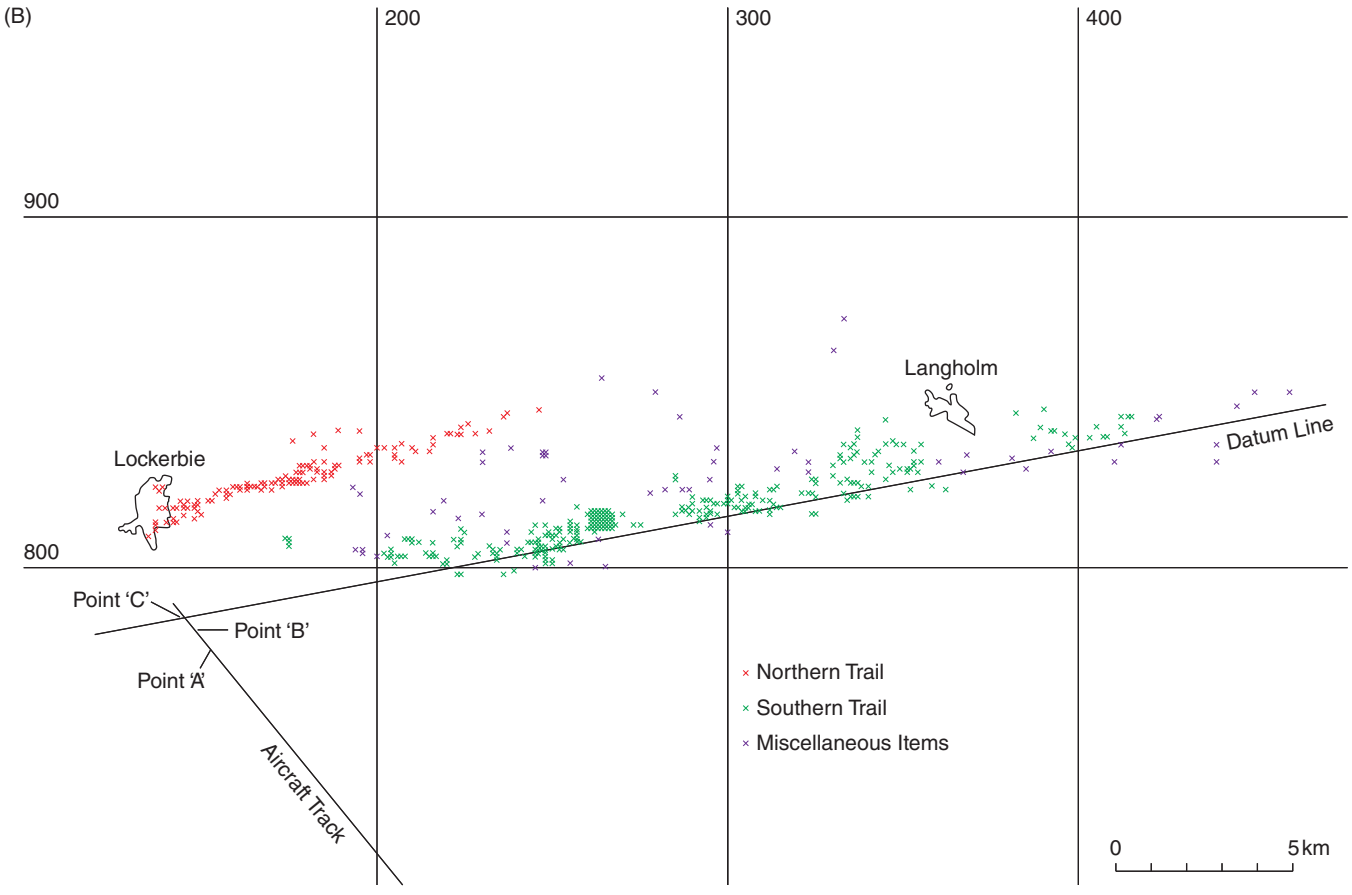


Figure 17.4 (*continued*) (B) Map of the debris trail; grid squares shown are 10 km in size. Detailed and up-to-date rural mapping facilitated collection of evidence from which the cause of the event was deduced. (Source: UK Air Accidents Board. Crown Copyright. Open Government License v2.0)

arranged in rows and columns (where each row relates to a discrete physical or human-defined object and the columns contain attributes of that entity) or as "spot samples" of a continuous field (Section 3.5.2). This is in reality a gross simplification; landforms are not always continuous, sometimes containing discontinuities or cliffs. The extent and naming of area objects is often inherently ambiguous (Chapter 4) and sometimes highly contentious (for example, the boundaries of disputed places such as Kashmir; see also Box 12.1); wars have often been triggered by such disputes.

The reality is that GI now exists in a growing number of different forms, each with different characteristics. Set out in Table 17.2 is a classification that we use in considering how GI is used in various domains. The first two rows are "traditional geographic information" dealt with in more detail earlier in this book; the others are more novel but increasingly of real significance. "Big Data" (Box 17.2) is often seen as encapsulating such diverse data types.

We obviously need to ensure our tools can deal with different forms of GI, but we also need to understand the implications for their use. As will become obvious, the trade-offs between the variables that may be collected, the accuracy, the spatial granularity, and the currency of GI change when administrative data and VGI replace traditional data collection mechanisms like surveys.

Trade-offs between variables collected, accuracy, spatial granularity, and currency change when administrative data and VGI replace surveys.

It is important to understand that transformations are often applied to a number of the data types in Table 17.2 to convert them into another type, as discussed in Section 12.3. This facilitates their visualization or comparison with other data. Thus it is relatively easy to convert suitably georeferenced lists of the locations of the homes of human individuals into aggregate data in field

Table 17.2 Classification of GI for the purposes of Chapter 17. Note that all categories below may be produced by official sources or commercial enterprises or provided by volunteers. Note also Section 2.4 on sampling, and Chapter 3 on representation.

Type	Subtype	Example(s)	Comments
Aggregate (1)	Continuous fields	Landscape/digital elevation model, remote sensing imagery	Usually sampled as array of heights or reflectance values or patchwise formulae (implicit x and y)
Aggregate (2)	Collection of individuals in area aggregate, each area considered as a discrete object	People in a given area	May be converted into a field, e.g., a density function
Lists of individuals	Individuals (people, fauna, etc.) considered as discrete objects	Each in form of (x,y,z_1 to z_n), e.g., characteristics and locations of human individuals, dwellings, or fauna at defined times	Can sometimes be highly sensitive because of privacy or national security considerations (Chapter 18)
Ambient and remote sensor data		Sensor information integrated to enable safe use of driverless cars	Key sensors involve measuring proximity to other vehicles, position on the road, etc.
Photographs of places: geopics		Technically similar to aerial or satellite imagery, but geometry often more complex and geography may be implicit rather than explicitly defined,	IARPA research project to geotag predigital imagery (Section 17.3.2)
Geovideos		e.g., Street View	Mostly just used for visual inspection? Limited use if these are out of date. Locally sourced street imagery is widely used in real-estate marketing and for monitoring the safety and insurance risk of closed premises
Verbal descriptors, including geography		Free text, e.g., certain historical novels, diaries	Extracting geographic location and descriptions of places into more conventionally structured form handled by GI systems can be tricky
Aural geographies		Spoken descriptors of place or people in the place; may well be in local language	Cultural-specific meanings may well be embedded. Can be very important to military activities.

Technical Box 17.2

Big Data and Science

The exponential growth in data collected has already been described (see Chapter 1). This has been going on for some time. Librarians and others have argued since at least 1944 that data volumes would outstrip capacity to store them—based on the observation that U.S. university libraries were then doubling in size every 16 years. The shift from analog storage—99% of all storage capacity in 1986—to digital (94% in 2007) has transformed the situation, at least temporarily.

There are many different definitions of Big Data, but the most widely used one is summarized by the 3 Vs—data volume, data velocity, and data variety. The first is simply defined as being too large to handle by standard contemporary analytical tools; the second is about how fast data are being collected; and the third is a way of describing the many different forms of data which are used—structured and unstructured (the majority), which are held in different types of databases as text documents, emails, imagery, videos, and much else.

It is clear that the definition of "Big Data" is subjective and relational, that is, what are Big Data will differ from one organization and perhaps even one sector such as multinational science to another such as local government.

Problem solving is all about separating the signal from the noise in data, usually statistically (see Further Reading). The essence of the case for Big Data is that data are now widely available, computer power is cheap, it is readily possible to hunt for associations between variables, and these give clues to the signal. Indeed Mayor-Schönberger and Cukier have claimed that:

- We can usually dispense with the need for new sample surveys (the cornerstone of most official statistics) given that we now often have access to huge existing volumes of data.
- Causality is less important than easily computed correlations based on large data volumes (from these we can predict the future provided we recompute frequently to cope with any change in the underlying relationships between variables; see also Chapter 1).
- We must get used to accepting "messiness" rather than expecting or searching for "privileged exactitude" from our data.

Much of this is uncomfortable for classical analysts brought up on the importance of sampling and identifying and understanding possible causal links. Fortunately it is also exaggerated and even misleading. Many Big Data sets do not cover entire populations (e.g., in the UK health retailers issue loyalty cards mostly to female customers) and hence the data are not representative of the whole population. It is also nonsense that significant imprecision can always be tolerated. In applications such as allocating resources to poor families, precisely matching what the law states is essential. Nevertheless, proponents of Big Data raise an important question in asking, "How good is good enough?" Answering the question is rarely easy—especially with GI—and ultimately involves judgment founded on a very clear idea of the users' needs, intimate understanding of the characteristics of the data, professionalism, and a code of ethics (see Section 18.5.2).

Finally, Big Data and Open Data only partially overlap. The first is predicated upon data volume, complexity, and massive analyses. The second is characterized by ease of access and reuse of data without significant penalties or difficulty. Open Data (see Section 17.4) can exist and be valuable without being voluminous. Many Big Data are held by private-sector firms like Google, Amazon, and satellite imagery companies, but the public sector—especially big science experiments—also hold and analyze colossal data volumes including GI pertaining to the 510 million km^2 of the Earth's surface and the 7 billion people living upon it in 2013.

form (i.e., a variable with a value everywhere in two dimensions). Overlaying multiple aggregate area datasets, each with different sets of boundaries, or combining one with a field variable is another matter (Section 13.2.4). The mechanics of the first process are embedded in most GI systems but this requires a set of assumptions about the spatial characteristics of the measurements within each of the areas. In some population-related cases this is facilitated by knowledge of the location of urban and rural areas (Section 12.3.3). In all such cases, however, uncertainty is being introduced into the newly created data in addition to what existed in the original dataset.

The characteristics of "traditional GI types" have already been described earlier in this book (Chapter 2). Now we consider those data types where changes in technology, use, and policy make them attractive for particular uses and where new problems arise.

17.3.1 GI About Individuals

Area-based aggregate data are relatively easy to handle but suffer two related methodological problems: the Modifiable Areal Unit Problem (MAUP; see Section 5.4.3) and the Uncertain Geographic Context Problem (UGCoP). The first is where any analyses of data (for example, correlation structures between geographic variables) are affected by the particular zoning system used. The second is where the analytical results are affected by the extent of any mismatch between the reporting zones used and the neighborhoods in which the data subjects are operating and whose properties are influencing their actions. Census tracts and city blocks are, for

instance, convenient data-reporting geographies but do not necessarily reflect functioning neighborhoods. Dasymetric mapping (Section 12.3.3) seeks to reduce the latter mismatch but is normally based on simplistic constructs such as outlines of built areas.

For these and other reasons, where data about individuals is suitably geocoded and is available (e.g., not constrained by privacy considerations), it is much more valuable than area aggregate data (to which it may be readily converted). It can, for instance, be aggregated and mapped in multiple different zone systems to check the stability of the results. It is particularly valuable where it also records changes through time (such as changes of location of the individual and hence exposure to different environmental hazards; see Section 17.3.3.1).

There are three broadly different ways of collecting information about human individuals: through surveys to collect new information, by reuse of existing administrative data, or by summarizing volunteered information. We now review each of the subcategories.

17.3.1.1 Survey Approaches

National statistical institutes (NSIs) exist in almost all countries, and many have long collected much of their information by way of formal, carefully organized sample surveys. All such surveys produce estimates of the (unknown) true value and of their accuracy. Unfortunately some NSIs do not have strong independence from political pressure, and this can lead to biased results favorable to the government. The International Monetary Fund, for instance, has severely criticized the quality of the Argentinean inflation data. Just as serious, a widely observed characteristic of statistical surveys across the world is that response rates are falling as citizens become ever more loath to fill in government questionnaires. In the United States response rates on average have fallen by around 20% over two decades. Finally, high-quality surveys are expensive because they necessitate a large field force. For some purposes, such as polling, Internet-based surveys have proved valuable. But given all the above, it is no surprise that official bodies are taking increased interest in information already held for other purposes by government.

17.3.1.2 Administrative Data

Much information in contemporary society about individual people, and their actions and transactions, wealth, and relationships and about individual businesses is now collected through administrative systems. Detailed personal data exist in almost all governments (and in many commercial organizations). In many cases the individuals have an incentive to report any changes (e.g., if they have another child and this entitles them to additional benefits). Such detailed information is now increasingly aggregated to produce information used in GI systems.

As usual, there is a trade-off here. The actual or potential loss of privacy (see Section 18.4) through misuse or loss of personal data has to be traded against the considerable benefits to be gained from use of individual data. Potential benefits include the following:

- Reducing the burden on people or businesses to fill in multiple questionnaires, with data collected for one purpose being spun off from databases originally collected for another.
- Knowing where everyone is based and hence the location of potential victims in the event of natural hazards occurring.
- Ensuring that wherever an individual travels, his or her detailed health records could be available to any doctor in the event of a sudden illness or accident.
- Allocating resources (e.g., social security payments) on a fair basis related to the individual's characteristics.
- Reducing the incidence of fraud by comparing living standards and so on of each individual with the norm for people or businesses of their type, often by merging multiple administrative datasets together (e.g., tax records and social security benefits).
- Tracking the life histories and geographies of individuals to study correlations between, for example, exposure to environmental hazards and subsequent illnesses.
- Profiling people on the basis of their background, personal characteristics, or contacts as being predisposed toward acts of crime or terrorism or for marketing or credit-rating purposes.
- Studying inequality in society through analysis of life experiences by people in small, specifically defined groups (e.g., communities or ethnic groups) from which new policies and actions might flow.

A geographic (or other) code of some kind attached to each record is required to produce the aggregate information (when needed) from the individual's details.

Given all the preceding points, the combination of GI systems and personal data brings many benefits for decision makers. Yet it is obvious that these benefits are bought at a price, at least for some people. The downside is potentially that:

- An individual's deeply valued privacy may be compromised.
- Fear of misuse of the data may undermine trust in the organization collecting it.
- Errors in data linkage could lead to incorrect judgments and policies.
- Administrative data are normally collected to meet specific purposes and hence are classified accordingly. Unsurprisingly, therefore, apparently similar datasets show different things: crime information is typically collected both from police recording systems and also via a sample survey of the total population. Box 17.3 illustrates this situation of multiple conflicting sources of information and what needs to be done to harmonize information from different administrative sources.
- The body collecting administrative data may decide to change what is collected to suit governmental policy purposes. This may well cause discontinuities in time series valued by other users.

17.3.1.3 Volunteered GI

We have already discussed VGI in various sections. Such information can assume many different forms, may be highly variable in quality and coverage, and may pertain to the physical or man-made environment or people themselves. It ranges from the often (but not invariably; see Box 12.1) astonishingly good quality of much Open Street Map (OSM) information to the partial and inconsistent. It can often be difficult to ascertain the quality of VGI even if published guidelines exist for those compiling it. Such data

Technical Box 17.3

Beyond a Traditional Census

Censuses of population have been held in the U.S. since 1790, in Britain since 1801, and in many other countries over periods of 100 years or more. The results of the 2011 UK Census were used to allocate the equivalent of many billions of dollars annually between local governments, health authorities, and other arms of the state—as well helping us to understand societal change.

The cost of that census was about some $650m, and the first results appeared over a year after the census date. Unsurprisingly, cheaper and faster means of producing the required data were demanded by politicians. To explore what was possible, the UK Office for National Statistics set up a major program to study the combined use of administrative datasets held elsewhere in government, individual data held by private-sector bodies (where those were available), and novel forms of survey. The challenges faced were numerous—highly technical, legal constraints on data sharing and political antipathy plus coping with public concerns about privacy and feared misuse of their data for detecting fraud, and so on.

The results are very specific to the national context. In countries with existing public registers (see Section 18.4), many of these challenges do not exist. For example, neither the UK nor the U.S.—unlike many mainland European countries—have a registration system whereby individuals must reregister with the state on moving to a different house.

Using the detailed 2011 Census data as a benchmark, different approaches were trialed. The results of using snapshots for the same date as the census from two major administrative databases—the number of people registered with a physician and those on the social security register—in raw form are compared in Figures 17.5A and 17.5B. The differences of the matches to the 2011 Census in different areas reflect the presence of military personnel in some large clusters, the low propensity of young males to register with physicians until they are ill, and other factors. The team subsequently devised methods of coping with these different counts, demonstrating a good ability to match census counts of population, though collecting certain attributes of the population would still require some form of survey. Based on this and other research the UK government has agreed that the 2021 Census will be a hybrid of Internet-based survey, some field survey, and some administrative data. The aim is to reduce the survey components over the years afterwards.

This approach differs from many of the commonly cited Big Data stories in recognizing conceptual and measurement differences between different administrative datasets and seeking to harmonize them to enhance the quality of the final result.

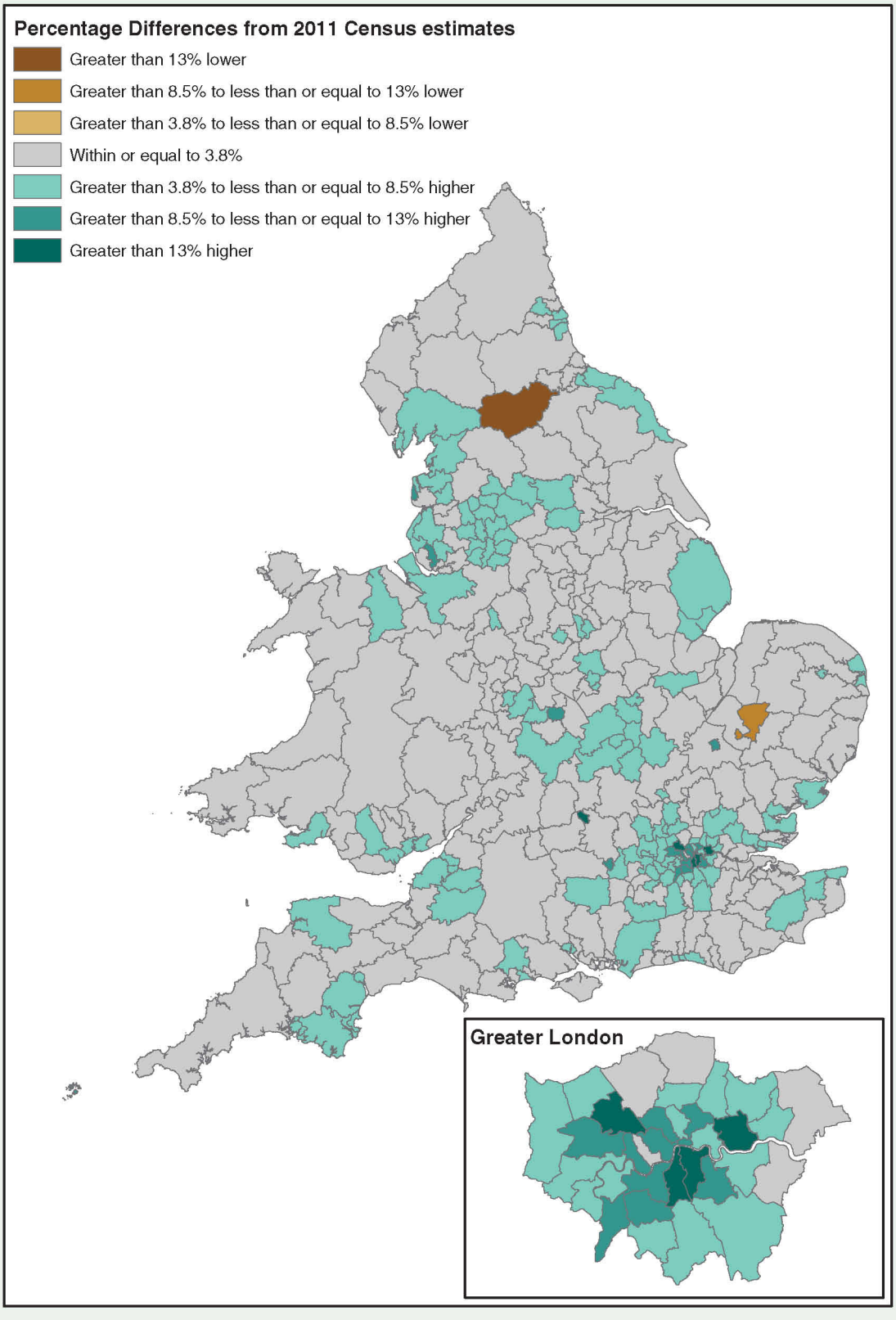


Figure 17.5A The geographic distribution of percentage differences between the 2011 Census population in local government areas in England and Wales and the totality of people registered with local physicians under the National Health Service on the same date.

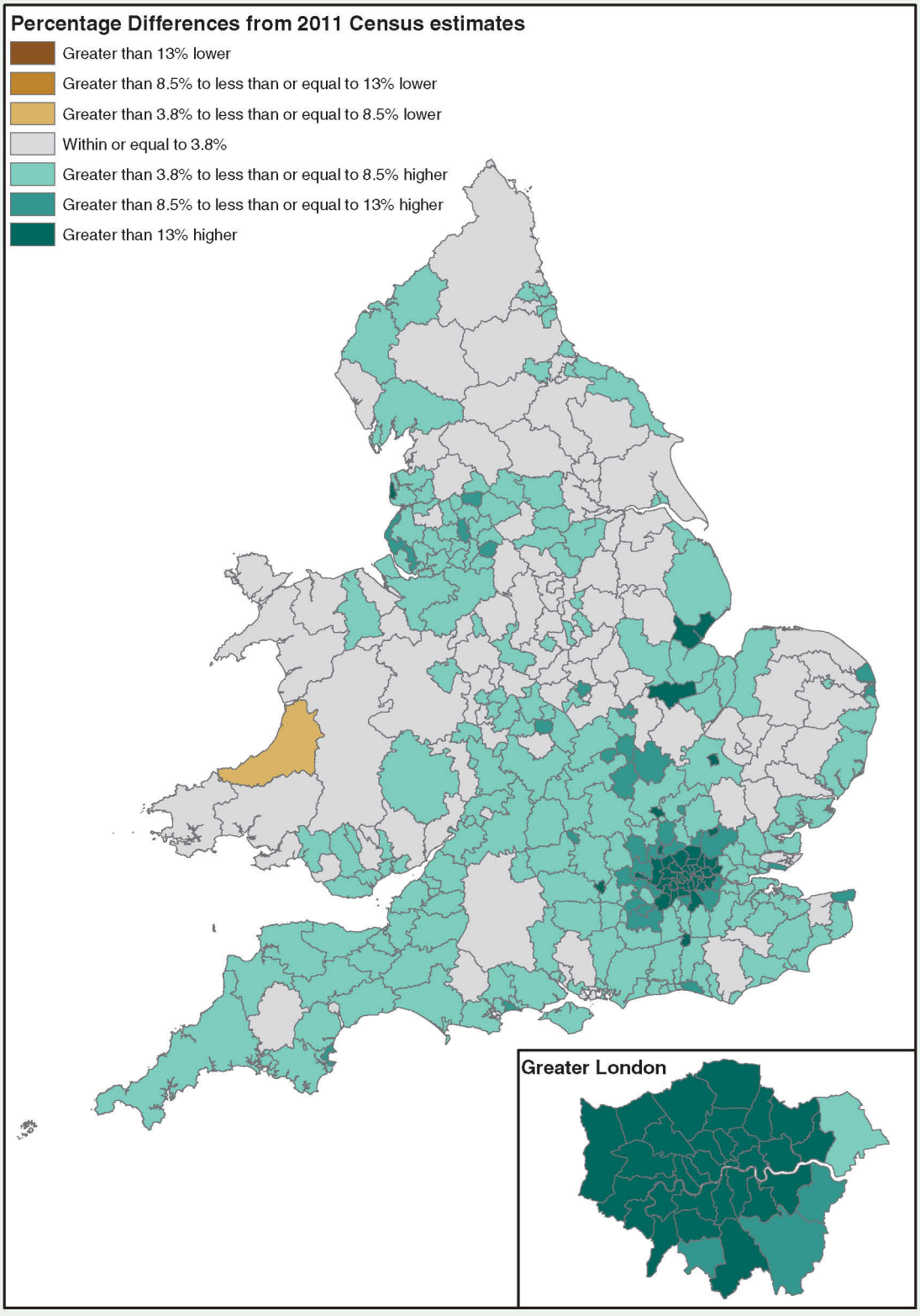


Figure 17.5B is an equivalent based on the numbers on the UK government's social security database at the same date. The gray areas are where the matches are within a tolerance of 3.8%. These are raw counts before modeling resulted in much closer matching.

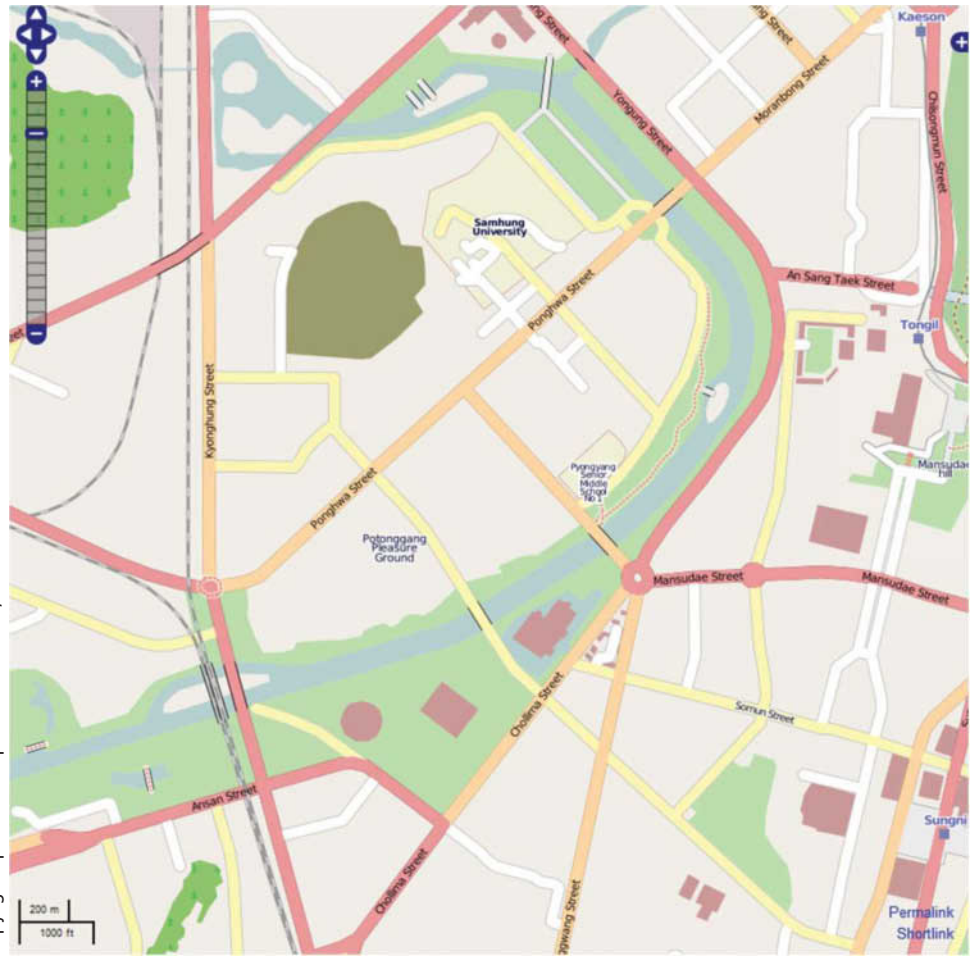


Figure 17.6 Open Street Map for part of Pyongyang, the capital of North Korea.

are often produced despite officialdom, rather than with its support or approval. Figure 17.6 illustrates a remarkable example of an OSM mapping of North Korea's capital city despite the regime's draconian security policies.

Here, however, we are mainly concerned with information about individuals, especially that volunteered about themselves. Perhaps the most active exploiter of high-frequency individual data is the retail industry (see Box 17.4). Individuals willingly provide personal data—some provided directly by them and much collected electronically from their purchasing habits—because of marketing incentives. Thus much use is made of store loyalty cards to analyze what many millions of individuals purchase, and where and when this is done. From this special offers are made, tailored to the individuals' interests, shopping habits, and activity patterns. The technology already exists to link mobile phone and credit card records of consenting individuals in order to monitor their movements and purchasing behavior—enormously valuable information to retailers. This technology and the information base held by Google and others enables geographically contingent marketing to be tailored to individuals, based on past purchasing preferences, similar profiles to other consumers, the location of friends, or other factors.

What makes this "volunteered information" is ultimately the law. Individuals agree to disclose their information and allow it to be aggregated for commercial purposes by the act of signing a contract. Usually this is achieved simply by ticking a box to accept the (often unread) many pages of terms and conditions. For this, the individuals gain access to services and incentives but trade off details about themselves that they would often resent giving to the government. This contractual commitment permits the service provider to exploit the information so long as it does not breach national privacy laws.

Increasing amounts of VGI are also derived from aggregations of highly disaggregated data obtained incidentally from individuals sending messages via Twitter, uploading photographs to Flickr, or simply having their cellphones on.

Application Box 17.4

Exploiting Customer Data

Created by its eponymous founders in 1989, dunnhumby (www.dunnhumby.com) is now owned by Tesco, the world's third-largest retailer. The firm also provides similar services to other major retailers in 28 countries (e.g., Macy's in the U.S.). Globalization of retailing is ensuring that many cutting-edge practices are spreading widely, and dunnhumby's business is expanding accordingly.

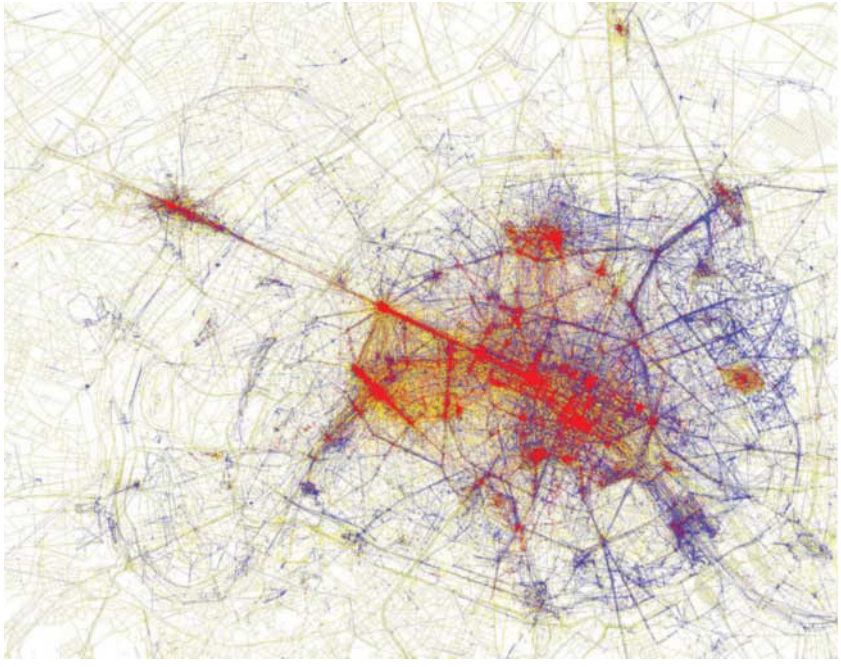
The main service dunnhumby provides is based on the data mining of information from sales, loyalty card use, and other data to enable retailers to understand customers and their needs, wants, and preferences. The scale of the mining, analysis, and exploitation is huge: Tesco, for example, has 15 million regular customers, and the geographical and other patterns of their purchasing are created from their shopping records and responsiveness to tactical marketing initiatives. More than 30,000 Tesco products are categorized individually to help build up a "Lifestyle DNA Profile" of each customer. On the basis of such analyses, each individual shopper is assigned to a group, and vouchers are printed for discounts on goods purchased in the past or for goods that other shoppers with similar characteristics have also purchased.

This results in a mailing to all 15 million Tesco Clubcard customers at least four times a year, with a summary of their rewards and vouchers tailored to encourage them to return and try new goods. Some seven million different variations of product offerings are made in each mailing. Customer take-up is between 20% and 50%, in contrast to the norm of about 2% in most direct marketing. The ranges of goods in store are also adjusted geographically in response to the habits of those who shop there. And the characteristics of new stores are planned on the basis of knowledge of people living nearby (including use of Census of Population and other externally provided data).

Figures 17.7A and 17.7B show fascinating spatial patterns at different scales. Such disaggregations can offer the potential for valuable insights into crowd behavior, the languages used by local residents, and even population migration (or at least local populations and those of visitors). The patterns, however, can be hard to interpret given the many interacting factors. Moreover for many purposes the data are only a numerator. Such raw data are dangerously seductive because they are often wholly unrepresentative of the underlying population: women, for example, are much more likely to have health store cards than males, and the great bulk of tweets are created by people between 20 and 40 years of age. To assess the real significance we often need at least to standardize such data with a denominator such as the total local population with cell phones or those using Twitter and geotagging their Tweets. Correcting for this bias may

Courtesy: Eric Fischer

Figure 17.7A The European distribution of geotagged Twitter and Flickr locations. Red dots are Flickr picture locations; blue dots are locations of Tweets, whereas white areas posted to both. The distribution reflects differing penetrations of Twitter and Flickr use, varied base populations, holiday travel, and other factors.



Source: created by Eric Fischer and made available from Flickr under a CC-BY-SA license. Base map is Open Street Map, available under a CC-BY-SA license

Figure 17.7B Visitors and locals in Paris. The latter (blue) are defined as not having taken any photographs in this area in the previous month, whereas visitors (red) congregate in tourist areas taking many pictures. Mixed areas are shown in yellow.

also require other information without the biases (e.g., official survey or administratively based information).

All this differs somewhat from crowdsourcing, where individuals select the cause to which they wish to contribute. This is now commonplace. One such GI example is the georeferencing of a variety of early historical maps, some dating back 400 years. Following the first public appeal by the British Library, some 700 maps from around the ancient world were georeferenced in less than a week using current mapping and the local knowledge of volunteers. Results are displayed using Google Earth to overlay the transformed map on contemporary mapping or imagery.

In some cases, however, involuntary information is collected without knowledge of the owner or data subjects. Mostly this is achieved by scraping Web sites to extract information (Section 17.3.3.2). Though in principle copyright restrictions exist (see Section 18.3.2.1.2), the reality is that once material—text, numbers, or images—has been posted on the Internet, subsequent control is limited.

17.3.1.4 Aggregate Data from Synthetic Individuals

The global coverage, quality, and currency of some crucial datasets—especially those pertaining to human beings at fine spatial granularity—are highly variable. Population censuses provide perhaps the best overall source but are held infrequently, and the variables summarized are often inconsistent between nations.

As a result, some enterprising organizations have sought to create information about the numbers of people distributed across the Earth by synthesizing it from a number of variables, some acting as proxies. The Landscan data product produced by the U.S. Oak Ridge National Laboratory GI system facility is a good example. This uses a mixture of data from many sources to model the global picture of ambient population (a 24-hour average) at 30 arc seconds resolution (c 1 km^2 at the equator). The input data include national midyear estimates of population published by the U.S. Bureau of Census, land cover, topographic mapping, administrative boundaries, and various forms of imagery data. The data are merged via a form of dasymetric mapping (Section 12.3.3). The Landscan team acknowledges that the multiple data sources vary considerably in characteristics between and within countries. So this may be the most consistent global information and the best population data that exists for some areas, but it is inevitably of highly variable accuracy.

A highly innovative example of synthetic individualized data relates to public health. In 2009, a new strain of flu H1N1 appeared in Asia. Health authorities were concerned that this would become a pandemic (see Section 19.6.3.2). At the time there was no effective known vaccine. A critical piece of information needed to take whatever action was possible was where the flu had spread; the sooner this was available the better. Typically quality-assured official statistics derived from reports by doctors take weeks to be available. In the United States, Google took the 50 million most common search terms in the three billion search queries they received each day. They then correlated the changing incidence of these search terms with the Centers for Disease Control's data files on the spread of seasonal flu between 2003 and 2008. They found that a small number of these search terms correlated with the

spread of flu as recorded in official figures in 2007 and 2008. Using this "the past is the guide to the present" approach, they were able to describe the spread of flu in near real time. This begs the question of how good was the proxy? Was it good enough for the purpose? And will the past always be a good guide to the present? Part of the answer was provided in a paper in *Science* in March 2014, which showed that the Google Flu Trends tool was overestimating flu cases by 30% or more when calibrated against flu reports provided to the U.S. Centers for Disease Control and Prevention (CDC) by doctors (which had a 2-week delay). The cause seems to have been a combination of human factors—people searching for information when they only had flu or because of a "snowballing effect" on hearing others had found apparently good guidance—and methodological shortcomings of the algorithm. The overall conclusion seems to be that such approaches are best regarded as exploratory and that there is value in comingling Google trends and official data.

17.3.2 More Novel Forms of GI

The BBC Domesday Project of 1986 (Box 17.5) was a conscious attempt to capture details of the whole UK 900 years after the original Domesday Survey. It involved bringing together government data, crowdsourced text and photographs, sounds, maps and imagery, and much else, and linked these on various criteria, especially location. It was the precursor for many contemporary Web-based systems.

Ever since Domesday, the forms of GI we seek to exploit have been multiplying. Table 17.2 highlights some of these. But the handling of some of these for certain tasks is often still difficult. For example, manual extraction of geographic characteristics from artistic images and written or (especially) spoken text is quite common, especially in regard to historical information. But automated extraction of such information for use in standard GI systems is difficult. Indeed, we have no agreed-upon tools and procedures for such automated extraction of "facts" from qualitative information, not least because this may be strongly influenced by cultural factors and language constructs (e.g., the huge range of words in the Inuit language for forms of snow). Potential applications of such GI are, however, widespread, including archaeology, historical comparisons, and in military intelligence. For instance the U.S. Intelligence Advanced Research Products Activity (IARPA) has advertised funding for research to find better ways to geotag predigital images.

17.3.3 The Changing World of GI

Here we highlight two major changes in recent years, over and beyond the growth of data about individuals.

17.3.3.1 The Rise of Geotemporal Data in GI

Some datasets are collected much more frequently than others; for example, meteorological data are generally recorded much more often than people are counted. Such data introduce temporal as well as spatial sampling issues (see Section 15.1). Thus Figure 17.2 shows a rapidly mutating radioactive plume emanating from Chernobyl.

Application Box 17.5

The Domesday Project

In November 1986 the BBC marked the 900th anniversary of the original Domesday Survey by William the Conqueror by carrying out a multisource survey of Britain. The result was arguably the first personal GI system. The BBC and its partners brought together 54,000 images (maps, photographs, satellite images), 300 mB of compressed official statistical data, and millions of words of text about 4 km by 3 km areas provided by one million members of the public—an early example of crowdsourcing. Included in the data were 21,000 files of spatial data showing national coverage down, in some cases, to 1 km^2 granularity. This data included geology, soils, geochemistry, population, employment and unemployment, agriculture, and land use/land cover. The different types of data were cross referenced by location or theme. Access to the data was by pointing on maps, inserting place-names or coordinates, or through use of a thesaurus. The file structures and location of the data were invisible to the user.

The platform used was the BBC microcomputer, which was then almost ubiquitous in British schools. The data were stored on a Philips Laser-Vision ROM. At the time the government's pricing policy was to charge for much of official data (Section 18.3.2.3.1). To have purchased all the official data held on Domesday would have cost about $375,000; the charge for the whole system to schools was $4500.

Domesday was essentially an early information infrastructure encapsulated in a box. Developments of the Internet and Web plus the huge increase in information available in digital form distributed across a multiplicity of sources rendered it obsolete. But it pioneered many of the concepts that we now take for granted.

In times past we have typically been content with—or at least had no option but to use—GI collected infrequently even if sometimes at fine spatial granularity. Thus most countries have traditionally updated their topographic mapping infrequently and collected detailed information about their population most commonly every 10 years. In some Western countries some of the population results are made available for hundreds of thousands or more small areas (such as city blocks). But this trade-off between space and time is now changing as new technology, changing user needs, and the public's growing unwillingness to fill in government forms evolves. Both official administrative data and volunteered GI provide the opportunity for increased frequency of information and greater analysis of geotemporal trends.

The most striking of these changes relates to transport data. Thanks to the ubiquity of smartphone apps and GPS, it is now routine in many jurisdictions to have access not only to public transport timetables but also to the actual location of the next bus, train, or plane in real time. In the U.S. and UK drivers may be charged for their insurance based upon where, when, and how they drive, computed using GPS in the car. This provides an incentive for careful driving. Nor is transport data only useful in individualized form. Computing speed of car movement from individual phone locations over time and aggregating these figures to provide advice to others approaching the same road gives a proxy for traffic congestion.

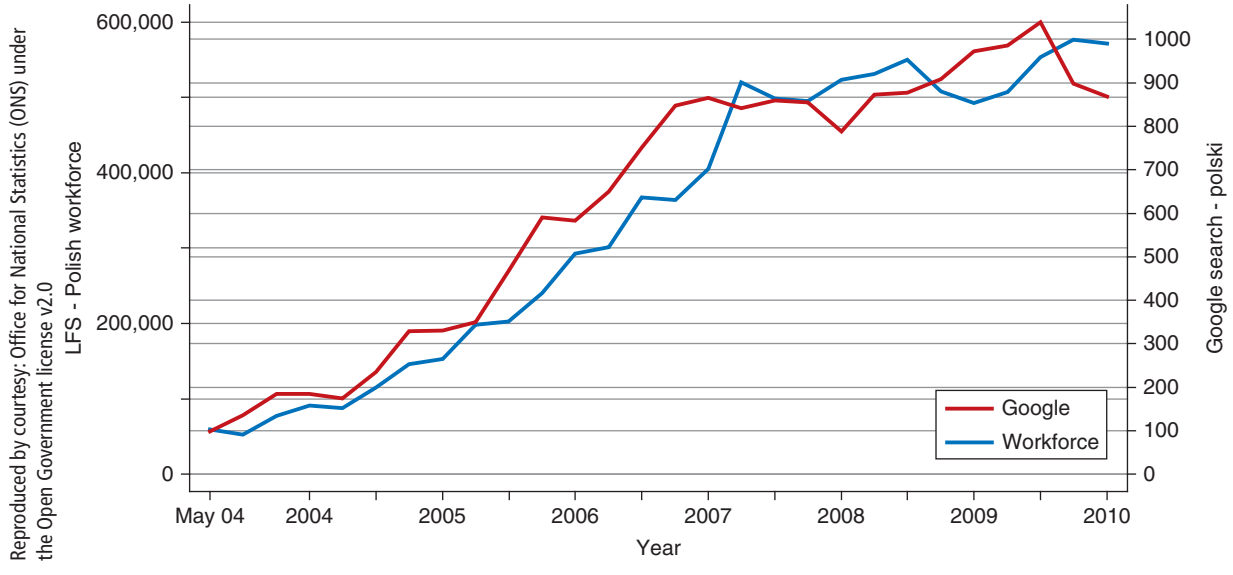
More directly individual is the use of GPS-enabled tags, often on the ankles, of those recently released conditionally from jail. This enables officials to ensure these potential recidivists do not go to areas from which they are excluded and to correlate their location and time with the occurrences of newly committed crimes. For different reasons, tags may also be fitted to those who are suffering from mental illness or some other debilitating condition such as dementia. Here the aim is to protect them from becoming lost.

17.3.3.2 Private or Public Sector?

Until recently, most of the benefits of GI arose from government collecting, holding, and analyzing raw data. But the situation has changed dramatically. For example, using Web scraping technologies to monitor online prices every day, PriceStats—a spin-off from MIT research—collects consumer price information from about 900 retailers across more than 70 countries. It publishes daily inflation series for over 20 countries, including both developed and developing economies. The organization offers a commercial service and claims that the trends in its data are very similar to those of more conventional data series. In effect it is seeking to provide a substitute for traditional sources of data.

One area where the private sector currently leads is in "nowcasting." This involves defining the current situation some weeks or months before official data are produced. The first papers to suggest that Web search data might be useful in forecasting economic and other statistics were published in 2005. The technique, mostly employing Google Trends, has since been used in epidemiology (Section 17.3.1.4), consumer sentiment, retail sales, and housing applications. Figure 17.8 shows a comparison between two

Figure 17.8 Comparison of the numbers of Polish people migrating to England between 2004 and 2011 derived from Google Trends and from the UK Labour Force Survey.



measures of migration—one based on conventional surveys, the other based on Google Trends.

All this raises a question of whether private-sector information is better able to describe many of the characteristics of human individuals and societies than traditional official statistics or other public-sector information. The answer is that it all depends—and is mutating. Certainly the social media tools provided by the private sector are now being used by billions worldwide. The information assembled using these tools is proving hugely valuable to the commercial sector. It is also of considerable interest to security organizations seeking to identify possible terrorists and to armed forces operating in foreign terrain.

But still greater benefits could certainly be realized if public- and private-sector information could be merged or comingled. The ability to merge customer preferences (e.g., for food, alcohol, or cigarettes) or spending patterns with health information could provide many benefits to the treatment of disease. Such cross-sector exchanges of data are fraught with difficulties given national legislation, commercial confidentiality, the nonrepresentative sample coverage of much commercial data, and the possible reactions of citizens and customers.

17.4 Open Data and Open Government

In Chapter 1 we introduced the concept of Open Data (OD); here we examine it as well as the drivers that have led to it becoming important in GI and far beyond in more detail. As ever, definitions vary between different organizations and authors. A simple description of Open Data, however, is "Data that can be freely accessed, used, reused, and redistributed without hindrance by anyone—subject only at most to the requirement for attribution and share-alike." Some bodies say that there may be some charge for OD, usually no more than the cost of reproduction.

Predicted benefits of Open Data—and prime drivers for political action—have included enhanced transparency and government accountability to the electorate, improving public services, better decision making based on sound evidence, and enhancing the country's competitiveness in the information business. The common finding worldwide is that GI is the cornerstone of success in making government information widely useful.

The U.S. federal government has always taken the view that the great bulk of information that it holds—other than that restricted for security or environmental protection purposes—should be made available at marginal cost or less and free of copyright restrictions (Section 18.3.2.1.2). In fact, that country pioneered the concept of a national spatial data infrastructure (Section 18.6.1), in many respects a partial precursor of national information infrastructures now under discussion. President Obama launched an Executive Order authorizing Open Data on his first day in office in 2009; the number of OD sets has risen from 47 in March 2009 to over 156,000 in late 2014.

The global policy picture has always been much more complex and variegated but has changed very rapidly in some countries since about 2007. The most evident manifestation of this is the growth of the Open Data and Open Government movements. In Britain a series of bodies to represent the user needs in relation to additional datasets and to carry out research and entrepreneurial activities with start-ups was set up. By late 2014 nearly 20,000 datasets had been made available in the UK as Open Data under a standard, simple Open Government License. These included many datasets useful to GI system or service practitioners such as numerous mapping files produced by the national mapping organization and real-time (e.g., transport) information.

Related and contemporary developments have also taken place in many other countries. Perhaps the most obvious demonstration of the internationalization of the Open Data movement is the charter signed by the leaders of Canada, France, Germany, Italy, Japan, Russia, the United Kingdom, and the United States at the G8 meeting in June 2013. This commits those nations to a set of principles ("Open by default," "Usable by all," etc.), to a set of best practices, including metadata provision, and national action plans with progress to be reported publicly and annually. Other global bodies—notably the World Bank and the UN Economic Statistics Directorate plus some U.S. states and cities around the world—have also committed to the principles of Open Data and enhanced access to their data stores.

These developments have forced debates about the extent to which government departments can be compelled to make available information at a time of austerity, the extent to which a national information infrastructure and strategy is required, and the trade-offs between making available very "raw data" quickly with little quality assurance (described by some as "fly tipping," i.e., dumping trash illegally) or slower delivery and better QA and metadata. It is now widely accepted that there is also a shortage of data scientists who are able to manipulate such Open Data but also understand its domain-specific characteristics.

A related development is the Open Government Partnership (OGP). This is an international organization with some 55 countries founded in September 2011. The objective is to promote transparency, increase civic participation, fight corruption, and harness new

technologies to make government more open, effective, and accountable. OGP is overseen by a multinational steering committee of governments and civil society organizations. To become a member of OGP, participating countries must embrace a high-level Open Government Declaration, deliver a country action plan developed with public consultation, and commit to independent reporting on their progress going forward. It is evident that progress can only be made through the use of widely accessible data—hence the link to Open Data and GI in particular.

17.4.1 The Metadata Issue

Throughout this chapter we have referred frequently to the quality and other characteristics of GI. Yet getting sound, consistent and user-focused metadata describing such characteristics has long been a problem. Previous schemes produced by government bodies have had limited success. Typically costs fell on the producers, whereas any benefits accrued to the users. One more recent Web-based scheme is the data certification mechanism proposed by the Open Data Institute.

One element of metadata concerns OD interoperability and avoidance of use of proprietary tools. Sir Tim Berners-Lee, founder of the World Wide Web, has argued strongly that all OD should be migrated toward more sophisticated formats to facilitate more effective use (Chapter 1). He devised a star-based rating system for summarizing the utility of Open Data, which is now in use in many public-sector organizations.

- ★ One star means the data are accessible on the Web. They are readable by the human eye, but not by a software agent, because they are in a "closed" document format such as PDF.
- ★★ Two stars mean that the data are accessible on the Web in a structured, machine-readable format. Thus, the reuser can process, export, and publish the data easily, still depending, however, on proprietary software like Word or Excel.
- ★★★ Three stars mean that reusers will no longer need to rely on proprietary software (e.g., use CSV instead of Excel data). Accordingly, reusers can manipulate the data without being confined to a particular software producer.
- ★★★★ Four stars mean that the data are now *in* the Web as opposed to *on* the Web through the use of a URI, a Uniform Resource Identifier. As a URI is completely unique, it gives a fine-granular control over the data, allowing for things like bookmarking and linking. An example would be an RDFa file containing URIs.
- ★★★★★ Five stars mean that the data are not only in the Web but are also linked to other data, fully exploiting the Web's network effects. Through this interlinking, data get interconnected whereby the value increases exponentially because they become discoverable from other sources and are given a context (e.g., through links to Wikipedia). An example would be an RDFa file containing URIs and semantic properties (allowing for linked data reuse).

It is relatively easy to imagine a corresponding ordinal scale describing the geographic characteristics of spatial data, such as the form of geocoding used.

Biographical Box 17.6

Sir Tim Berners-Lee—Web founder and Open Data champion

A physics graduate of Oxford University, Tim Berners-Lee (Figure 17.9) invented the World Wide Web, an Internet-based hypermedia initiative for global information sharing while at CERN, the European Organization for Nuclear Research, in 1989. He wrote the first Web client and server in 1990. His specifications of URIs, HTTP, and HTML were refined as Web technology spread.

He is the 3Com Founders Professor of Engineering in the School of Engineering with a joint appointment in the Department of Electrical Engineering and Computer Science at the Laboratory for Computer Science and Artificial Intelligence (CSAIL) at the Massachusetts Institute of Technology (MIT). He also heads the MIT

Source: World Wide Web Consortium

Figure 17.9 Sir Tim Berners-Lee.

Decentralized Information Group (DIG). He is also a Professor in the Electronics and Computer Science Department at the University of Southampton, UK.

Tim is the Director of the World Wide Web Consortium (W3C), a Web standards organization founded in 1994 that develops interoperable technologies (specifications, guidelines, software, and tools) for the Web. He is also a Director of the World Wide Web Foundation, launched in 2009 to coordinate efforts to further the potential of the Web to benefit humanity.

He has promoted Open (government) Data globally, persuading two successive British Prime Ministers (and many other senior players worldwide) to become its advocates on the basis of enhancing transparency, good governance, and the creation of new technology-based industries by entrepreneurs. Tim is president of London's Open Data Institute. Among the many awards that he has received was the newly founded Queen Elizabeth II $1.5m prize in 2013 in Engineering for "ground-breaking innovation in engineering that has been of global benefit to humanity." He shared this with Vinton Cerf, Robert Kahn, Louis Pouzin, and Marc Andreesen, fellow pioneers of the Internet or Web.

Source: Berners-Lee biography at www.w3.org/People/Berners-Lee/

17.5 Example of an Information Infrastructure: The Military

Thus far we have talked in the abstract or given short examples of how an information infrastructure is crucial to organizations or governments. We now give a more concrete example.

We can trace the beginnings of much national mapping to the needs of the military. This is not surprising: the first duty of the state is to protect its citizens and their interests. Sometimes such defense of national interests also involves operations far from home. Unsurprisingly, therefore, many countries maintain archives of current and historical maps of areas outside their own terrain. The Soviet Union had an extensive mapping program outside its own borders until its collapse. Effective defense necessitates the state having the wherewithal to make such protection effective, including a decision-making apparatus and an information infrastructure. Decisions, often based on the best available geographic information, are made from top to bottom of every military organization every single day.

This section is based on information in the public domain and relates only to developments in Western countries. It excludes any consideration of the substantial GI and GI technology activities of the security services, such as the U.S. National Geospatial-Intelligence Agency (NGA), which has overall responsibility for GEOINT (see Section 17.5.2) in the U.S. intelligence community. We can also be certain that the military information infrastructures in many countries are, together with those of the security services, by far the most extensive, sophisticated, multinational, and integrated of any in the world.

17.5.1 Technological Change and the Military

The most familiar role of the military is in warfare. There have been around 200 wars since 1945. Technology has long been influential in reshaping intelligence gathering and the way warfare is executed. It has radically changed mechanization and communications; the evolution of air power since 1918 has been a particularly important factor. In spatial terms these developments have collapsed distance strategically, operationally, and tactically. Now the entire world is under surveillance from satellites and some of it from unmanned autonomous vehicles (or UAVs; see Figure 17.10); armed drones are guided from command centers thousands of miles away. Elite infantry and armored units can be transported quickly by air and deployed in a matter of a few days yet be in constant touch with headquarters.

The changes have been so profound that some historians have argued that geographic space has ceased to be an encumbrance, let alone a friction. This extreme view ignores the reality of operating in hostile terrain, with equipment that does not always work as intended. Moreover the increasing need to engage verbally with local communities has been a feature of recent conflicts since tank warfare across the European plains has given way to urban and guerrilla warfare. Social science and mapping of community characteristics have become increasingly used to help military personnel understand, differentiate, and engage successfully with local populations.

Finally, underpinning all the obvious military tasks is a variety of other roles that require information of different types and latency. For example, estate

Source: (A) estt/iStockphoto; (B) Richard Watt/UK Ministry of Defence under the Open Government License 2.0; (C) U.S. Geological Survey

Figure 17.10 A to C Examples of Unmanned (flying) Autonomous Vehicles used for imaging or other surveillance purposes. Some of these are now increasingly being used for a variety of civilian purposes as well as military ones.

management and logistics are of huge consequence. Running many military bases requires management of a complex array of resources, assets, facilities, lands, and services similar to those of a medium-sized city—and using GI systems in the same way as in the civilian environment. Getting stores (including ammunition) to the right area when needed is similar in principle to the role of transport logistics companies such as FedEx but complicated by the hostile environment in which the military sometimes has to operate.

17.5.2 The Military Information Infrastructure

Success in operationalizing the concept of command, control, communication, and coordination in military operations is substantially dependent on the availability of accurate GI, typically integrated from multiple sources, to arrive at quick operational decisions. Though interoperability is crucial, it is also essential that it does not extend to any enemy forces! Table 17.3 sets out a very simple summary

Table 17.3 Levels of decision making in the military and information required.

Level of decision making	Information required
Grand strategic	Overviews suitable for discussions with politicians (e.g., Common Operating Picture graphics)
Strategic	Synoptic views of assets and challenges to them plus plans for action (e.g., battle theatre plans)
Operational	Information needed by one-star general leading the activities of a regiment or brigade (battle scenario maps and graphics)
Tactical	Detailed information on the local area needed by front-line troops (e.g., tactical maps)

of the types of geographic information needed by Western forces for different purposes.

As in other application areas, GI systems allow the user to capture, manage, analyze, visualize, and exploit geographically referenced information about physical features and much else. For instance in planning campaigns military field commanders would like to know terrain conditions, especially elevations, for maneuvering armored carriers and tanks and for use of various weapons. In addition, they need information on vegetation cover, road networks, and communication lines to deploy resources effectively. All these details must be available to field commanders in a coordinate system that matches with the equipment they use for position fixing. Any discrepancy in these inputs may endanger the operation and the lives of troops.

In military parlance, GI is often called geospatial intelligence, or GEOINT. This is normally defined as "information about any object—natural or man-made—that can be observed or referenced to the Earth and that has national security implications." It is derived from the full range of current and historical information sources described in earlier chapters.

In addition to GEOINT, signals intelligence (SIGINT), human intelligence (HUMINT), and open-source intelligence (OSINT) contribute to the military information infrastructure. These are most useful when geocoded and integrated with each other. SIGINT, for example, played a significant role in identifying the location of the leader of Al-Qaeda. HUMINT is intelligence derived from information collected and provided by human sources. Finally, OSINT is collected from publicly available sources, most typically now from analysis of material on the Internet, including social media. The integration of such a variety of information of very different characteristics and uncertainty is challenging but essential.

17.5.2.1 MGCP: Example of Multinational Collaboration

Military organizations are typically parts of a major alliance; NATO is the most well-known example. This means that additional resources may be brought to bear in times of stress, including access to mapping and GI.

A striking example of such collaboration is the Multinational Geospatial Co-Production Program (MGCP). This program had 29 nations as members in 2013. It involves the creation and maintenance of fine-resolution vector data at accuracies equivalent to 1:50,000-scale mapping. The basic data unit or "exchange unit" is a 1-degree cell. Marketing-style incentives are used to maximize inputs of units by the partners to the database; the numbers of units that can be extracted by a member country rises on a sliding scale as the number they input increases. After inputting above 200 units, the donor country may access all the units available provided they also take part in quality assurance. Such an arrangement reduces duplication of effort and has been demonstrated to work effectively in emergencies. For example, the UK's Defence Geographic Centre now has the capability to produce standard map sheets from the data to an internationally agreed specification and distribute them to its military customers in less than 20 hours.

Whereas all systems are now digital, different sections of the armed forces require different products including paper-based ones. Those in armored vehicles or aircraft have the capacity to use heads-up and other digital displays. But for the infantry, any extra weight to be carried, the danger of being located by light emission, and the possibility of digital jamming all ensure that paper maps are often favored, especially with topographic maps and imagery arranged back to back.

The range, magnitude, complexity, and interrelationships of information that is within scope for military personnel and the frequent need for speedy analyses is such that "Big Data" (see Box 17.2) is inevitably of great interest to the armed services. Yet large-scale data mining and delivery capabilities have some challenging consequences. For instance, the wider the access to information the greater the likelihood of leaks of confidential information (such as appeared on Wikileaks).

17.5.3 Civilian Spin-Offs

There are at least three areas where spin-offs of military capability benefit the civilian sector. These are the development of new technologies, availability of GI for civilian purposes, and the use of military personnel in civil emergencies.

The most influential military technology by far used in the civilian sector is, of course, the Global Positioning System, or GPS (see Table 1.4, Section 4.9, and Box 17.7). This has totally transformed the entire GI industry through changing the way humans—and machines—can describe and interact with geography.

Substantial taxpayer investment in military R&D led to the development of ARPANET, the precursor to the Internet. The success of the U.S. dominance in

Technical Box 17.7

The History of GPS

In the early 1960s the U.S. Department of Defense (DoD) began pursuing the idea of developing a global, all-weather, continuously available, highly accurate positioning and navigation system to meet the needs of a broad spectrum of military users. After much experimentation with different approaches, approval was given for full-scale development in August 1979. This was punctuated with problems: budget cuts led to the program being halted for two years.

The crucial demonstration of GPS's military value came in the 1990–1991 war in the Persian Gulf. It enabled coalition forces to navigate, maneuver, and fire with high accuracy in the extensive desert terrain almost 24 hours a day despite frequent sandstorms, few paved roads, no vegetative cover, and few natural landmarks. So effective was it that the DoD had to hastily procure more than 10,000 commercial units to meet field demands.

The first U.S. pronouncement regarding civil use of GPS came in 1983 following the downing of Korean Airlines Flight 007 after it strayed over territory of the Soviet Union. President Reagan announced that the Global Positioning System would be made available for international civil use once the system became fully operational. In 1991 the U.S. Federal Aviation Administration promised that GPS would be available free of charge to the international community beginning in 1993 on a continuous, worldwide basis for at least 10 years.

All this led to the expansion of the surveying market by the mid-1980s even while GPS was still in development and only a small number of operating GPS satellites were in orbit. Surveyors also pioneered some of the more advanced differential GPS techniques, such as kinematic surveying. This development generated commercial revenue for U.S. equipment manufacturers to invest in new developments—a positive feed-back mechanism. The later development into a ubiquitous and free service has transformed the lives of many of the world's people.

Source: *GPS history, chronology, and budgets* at www.cs.cmu.edu/~sensing-sensors/.../GPS_History-MR614.appb.pdf

commercial fine-resolution satellite imagery industry is due in part to the military being the guaranteed "anchor tenants." UAVs (Section 17.5.1 and Figure 17.10) are becoming very widely used in the civilian domain. The U.S. Geological Survey has used them for wildlife monitoring, checking fencing, and tracking fires; others have used them for collecting samples from volcanic eruptions, surveying, and much else.

Inherent in military operations is the need to create GI if that does not already exist. An early example of this was making publicly available the Digital Chart of the World in 1993. This was the first globally complete 1:1 million standard map series in computer form. Another example is the detailed 1:5000-scale mapping of all Helmand Province and of cities across Afghanistan. Such information (as well as physical) infrastructures created for war are potentially of great value in subsequent peacetime.

The military also provide support of and intimate interoperability with civilian agencies. Providing search-and-rescue missions after natural disasters is a common and vital role worldwide for military forces.

17.6 Conclusions

We have shown how understanding the characteristics of information, especially that of GI, is crucial to many everyday tasks. The situation is not stable: nontraditional forms of GI—such as detailed administrative or personal data—are evolving as technology and user needs mutate. Information from difference sources and with different specifications is being comingled, with implications for its safe use. More positively, many governments have come to realize the potential of Open Data and begun to commit to making it more of a reality: GI, especially core reference data, is central to its success. But success also depends upon the existence of enough suitably trained data scientists who combine analytical skills with domain knowledge. In particular, GI expertise is essential to make our visions a reality.

Questions for Further Study

1. Suppose you are in charge of a business providing GI services or you are an army general. What kind of information and software tools would you need to make effective decisions?
2. What are the main characteristics of information, and how do these differ from those of physical goods? Are there ways in which GI characteristics differ from those of other information?
3. What is a national information infrastructure? Does it already exist? And why is it important?
4. Search the Web for national policies and activities on Open Data and summarize their differences. Why are there differences in progress between countries?

FURTHER READING

Institut de Radioprotection et Sûreté Nucléaire (IRSN). 2011. *Chernobyl's accident.* See www.irsn.fr/EN/publications/thematic/chernobyl/Pages/overview.aspx and simulation.

Kolvoord, R. and Keranen, K. 2011. *Making Spatial Decisions Using GIS: A Workbook*. Redlands, CA: Esri Press.

Mayor-Schönberger, V. and Cukier, K. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*, London: John Murray.

Shapiro, C. and Varian, H. R. 1999. *Information Rules: A Strategic Guide to the Network Economy*. Cambridge, MA: Harvard University Press.

Silver, N. 2012. *The Signal and the Noise: The Art and Science of Prediction*, London: Penguin Press.

Issues of *Trajectory: The Official Magazine of the U.S. Geospatial Intelligence Foundation*. Herndon, VA: U.S. Geospatial Intelligence Foundation.