Sharing Geographic Data: How to Update Distributed or Replicated Data

Andrew U. Frank

(Dr. DI Andrew U. Frank, Geoinformation Technical University Vienna, Gusshausstraße 27-29, 1040 Wien, Austria, frank@geoinfo.tuwien.ac.at)

1 ABSTRACT

Geographic data is expensive to collect and maintain and sharing data is crucial for its effective use in urban planning at all levels. For a few hardly ever changing themes the simple distribution of copies of data is feasible, but for other data, access to "live" data and updating, sometimes even distributed updating, of the data is necessary.

The organization of sharing data can be separated into three sets of issues: (1) *Interpretation*: how to understand the data, (2) *Authorization*: is a user permitted to use the data, and (3) *Access*: how to achieve effective and non-disturbing use and updating of data by several users? Solutions must take threats into account: hackers may try to steal or disturb the use of data, and the revelations of Snowden's documents only emphasize the danger of others reading data not intended for their eyes.

Effective sharing geographic data without conflicts requires integrating results from different areas of computer science research, including at least: cryptography, computer security, database management, and computer networking.

2 INTRODUCTION

Geographic data is expensive to collect and maintain, and sharing data is crucial for its effective use in urban planning at all levels. To use data collected by others is not without problems: one needs arrangements to understand the encoding, contracts to permit the use of the data, and finally methods to access the data. To have others use data that one has collected is not easy either: the owner of the data must insure that the data is only used by the users he has authorized, and only for the purposes permitted; it is too easy for any user to copy data and pass it on to others. Given the sensitivity of the European public, if data thought to be "personal" and given to a public agency appears on servers without protection, it quickly causes a public scandal. It is therefore important for public agencies to insure that all aspects of protecting the data are enforced, while at the same time allowing for maximal use of public data by others a contradictory charge!

The sharing of data can be organized into three sets of issues: (1) Interpretation: how to understand the data, (2) Authorization: is a user permitted to use the data and (3) Access: how to achieve effective and non-disturbing use and updating of data by several users? Solutions must take threats into account: hackers may try to steal data or disturb the use of data, and the revelations of Snowden's documents only emphasize the danger of others reading data not intended for their eyes.

The effective sharing of geographic data without conflicts requires the integrating results from different areas of computer science research, including at least: cryptography, computer security, database management, and computer networking. The proper solution depends on the granularity of the data (the size of a data element that can be processed meaningfully), on network connectivity (high vs. low bandwidth, permanent vs. intermittent), on the frequency of updates, etc.. A layered approach of stacked protocols suitable for different situations is expected to allow adaptation to different needs.

Much of the complexity of organizing data sharing is caused by the unfortunate mixing of these three concerns. The technology available is sometimes offering similar technical services and comparable methods to control authorizations and access, but combining them quite differently. Further confusion is caused by results being described in alternative terminology. Part of the confusion is created by the multiple meanings of the same words, depending on whether the context is semantics, authorization, or technical access: the word "producer" for example may refer to the organization (the preferred meaning here); or it may refer to the person within the organization who collects and encodes the data. "Use" may refer to reading data only, or may include access to live data and changes to them. In a technical setting, the word "server" is used for the technical system (of the producer) where the data is stored, and "client" is used for the system where the data is used.

ISBN: 978-3-9503110-6-8 (CD-ROM); ISBN: 978-3-9503110-7-5 (Print)
Editors: Manfred SCHRENK, Vasily V. POPOVICH, Peter ZEILE, Pietro ELISEI

3 SEPARATION OF ISSUES REGARDING INTERPRETATION, AUTHORIZATION, AND ACCESS

The issues around sharing of data can be separated into three concerns:

Interpretation: does the data make sense in the context of the users?

Authorization: are users permitted to use the data?

Access: how to get the data?

In order to effectively share data, an arrangement on interpretation, authorization, and access must be found.

Sharing data between organizations confronts the meaning the producing organization (and in particular the person collecting the data) gives to the encoding, with an understanding of the codes used by the using organization. It is not required (and probably not feasible) to have the meanings correspond completely, but only that the differences in interpretation are not leading to errors in decision making.

Sharing data between organizations requires administrative arrangements in which the producing organization allows the using organization to use a set of data for certain tasks and in a certain modes, e.g. read or update. The arrangement must be made between the legal entities which produce respectively use the data, but bind persons acting on behalf of the organization as well; this includes the trivial case where the legal entity is a single person. The organizations can then use contractual and technical means to enforce the agreements and to ensure that all obligations are met. For example, map data provided by a cartography publisher is given with strict limits where and how often it can be used and, obviously, with the obligation not to make it available to others.

If the user is allowed or required to update data, then arrangements must cover accountability requirements of the original producer and keeper of the data; it is usually required that the name of the person changing the data as well as their function within the organization is included in the recording of the change.

Sharing data between organizations finally requires technical arrangements to transfer the data between the systems involved. The transport of media, where the data is stored, is a simple method; it implies arrangements to use compatible methods for the storage and encoding of the data. Data can be communicated through the internet, either the full data set at once or individual pieces as they are required, and feedback from the using organization can flow back via the same path.

4 THREATS

The data must be protected from dangers which threaten its long-term usability and which may compromise the confidentiality of the data. Threats can be differentiated according to the three topics above:

4.1 Changing Interpretation

The interpretation of natural language terms changes Fleck [1981] ; a geographic example is the changing definition of a habitat. Widely published was the change of the defintion of “planet” by the International Astronomic Union in 2006, which made Pluto not a planet anymore but a “dwarf planet” instead.

4.2 Unauthorized use

Unauthorized use of data is use by persons not authorized to use it or access by persons which are authorized in some organization, but try and access the data in an inorganizational environment that is not authorized. For example, a person could be authorized to access data for their job in a planing authority, but then accesses the data while working in another job, e.g. for a bank. Sometimes the producer of the data puts strict limits on what the data can be used for, and thus non-authorized use would also be any use for purposes other than the ones specifically permitted. For example, data about buildings may be available for planning, but its use for taxation may not be permitted. Often authorized use does not include making copies and carrying data outside of the permitted environment; famous examples are bank employees who make copies of lists of bank clients and their account details.

A specific threat is from a person masquerading as an authorized user and presenting the identity of this user to gain access to data. Unauthorized persons can learn the identity of authorized users by observing their actions, or by evasdroping on their communication with the system.

4.3 Failure of access

Data may be deleted and lost, making access impossible. But access can also be hindered temporarily by malfunctions in the software or hardware. Data sharing systems are often complex, and many softand hardware components have to cooperate properly in order for the sharing of data to work. Ordinary mishaps, failure of technical components, and accidents are frequent; human error (i.e. plain stupidity) is a major cause, but increasingly criminal actions against data are observed as well.

5 ACTIONS TO COUNTER THE THREATS

What are actions producers of data must take to protect their data? The actions to protect data can again be differentiated along the three issues listed above.

5.1 Document interpretation

Proper documentation of the intended meaning of codes is important; it is part of the metadata which can be described following accepted standards (eg. Weibel et al., 1998). The interpretation of common terms often differs between agencies, and may also drift over time. A well-known example is the evolution of terms describing habitats, which change with the progress of science and may be confused with real changes in the habitat [Comber et al., 2004]. The use of qualified names following the RDF standard [Manola et al., 2004] allows differentiation between two agencies' use of the same term, or connecting a term with its proper meaning according to the year of the definition, thus differentiating the meaning of a term in an earlier and a later defintion.

5.2 Check changes to preserve interpretation

Data can degrade if updates include errors in the data. To prevent this, changes by users are checked against rules which fix the interpretation of the data. A person must have a name, a building must have a number of floors, etc. At the end of each changing operation by a user, the new data is checked against these so-called consistency rules, and an update is only performed if the new data conforms to these rules.

Changes are recorded with the time and date of the change, the authorized person that entered the change, and finally some justification for the change, e.g. a reference to a document or a contract. These records of changes guarantee auditability, which means that all changes in the data can be traced back to an authorized person who can justify the change. This is obviously of prime importance in systems dealing with ownership of land, but it is equally important for maintaining restrictions in urban planning.

5.3 Check authorization

Authorization is the process of connecting a real person or organization with an identity within a computerized system. The most common form is a user name (associated with individual persons) and password, which demonstrate that a particular technical client is acting on behalf of the person with the given user name. This of course works only if the passwords are kept secret, and not written down on small post-it notes and pasted on the monitor but then again: who can remember all their passwords? And how often is it an assistant performing some action on behalf of their superior, requiring the password to be passed on, defeating the purpose of it?

Authorization can be organized better than with a simple username and password. One effective method is certification, where a trusted party signs credentials for another, which then signs credentials in turn. Credentials are electronic documents which are protected by cryptographic means against forgery. Authorized users are given credentials which are properly issued and presented to gain access. The technical solutions are such that credentials are never transmitted as clear text, so others cannot eavesdrop on them.

5.4 Protect data during transmission

Confidentiality is threatened during the transfer of data between the technical systems of the organizations involved. Safeguarding the transport of the medium on which the data is stored is the least expensive method (e.g. through transport by a trusted person). For transfer over the web, encryption is effective, easy, and inexpensive.

5.5 Prevent loss of data

Replicating data together with recorded logs of all changes guards against accidental loss of data due to technical or human failures. The current state of the data collection can be reconstructed from snapshots and the log: all changes performed since the shapshot was made are applied to reconstruct the last state before the loss of the data occurred.

5.6 Prevent loss of use and access to data

Even a temporary loss of access to data may cause problems for a client. Technical systems are often duplicated, so that in case of the failure of one system the duplicated system can take over. High-availability systems achieve nearly 100% continuous service at a cost: the shorter the maximum tolerated interval without service, the more technical effort is required to achieve it, and the higher the cost. The outage of a town planning system for several days may be painful, but tolerable; a system to keep track of the current positions of a taxi fleet will quickly incur additional costs during the time that it is not accessible, and thus duplication of some crucial components may be justified.

6 TECHNICAL CONSIDERATIONS

6.1 Semantics

The meaning of the data, i.e. how the facts reported are encoded, is often described in additional documents, formalized as metadata or in informal descriptions.

Essential to sharing data is an understanding of the data by the user: what does the code used mean? The producer gives meaning to the data according to his viewpoint and the aspects important for the intended use of the data collected. The user uses the data with a semantic schema, which must at least be coherent with the one used by the producer. Semantic coherence does not imply that the data has the same meaning, but only that the conclusions the user draws from the data are not in contradiction to the producer's interpretation.

6.2 Authorization

The legal agreement permitting a user to use a data set from organization must be mapped to the realm of data. Specifically, the user and the organization (short "legal entities") must be represented in the data realm. A "token", i.e. a small set of data, is uniquely assigned to each user so that no two users have the same token, and there is a method to confirm the association.

The construction of a personal token is trivial any random data with sufficiently small chance of accidental duplication is suitable. The association of a token with a legal entity is either through a hierarchy, or a network of trust. Organizations typically use a hierarchy and get a certificate signed by a trusted certification organization, which associates their token with their name (important is the U.S. company Verisign, which issues "root" signing certificates). Persons more often use a network of trust: other people sign associations between persons and their tokens, and users of the tokens can inspect who signed for them. [http://en.wikipedia.org/wiki/Pretty_Good_Privacy].

The method is based on public and private keys [http://en.wikipedia.org/wiki/Publickey_cryptography], so that each user can publish their public key (derived from their token) and must keep their private key secret. Other users can decode documents signed by a user with their private key (which only they know) and are then assured that the document really is from that user (e.g. an email sent to them). Other users can encode documents with the public key of another person and send them away; they are guaranteed that only the intended person can decode the document with their private key.

6.3 Digital Watermark

Authorized users of data may be tempted to give the data to others despite the fact that their authorization to use the data does not allow them to do so. To prevent such unintended distribution, which may cause commercial losses, the data can be watermarked [http://en.wikipedia.org/wiki/Digital_watermarking]. For each authorized user, invisible marks are included in the data before they are given to the user. If later the same data appears with an unauthorized user, the hidden digital watermark allows the copy to be traced back to the authorized user who has leaked the data in violation of their obligation to keep the data secret. For geographic data, digital watermarks have been proposed not only for image data but also for vector data; the

difficulty lies in hiding the marks within the data so that they are not detectable by others, and do not disappear through simple coordinate transformations [Ohbuchi et al., 2002].

6.4 Keeping data current

Most GIS data changes with time. Very few data sets are static and remain the same forever (not even digital terrain data). But how to distribute the changes?

Distribution of snapshots. A snapshot of updated data can be delivered periodically to the user; this limits how "out of date" the used data can become. Problems emerge if the users connect the data they receive with their own data (and not just graphically overlay them); the connection between the data received and the own data of the user is through some data elements which serve as identifiers these must not be changed by the provider from snapshot to snapshot.

Distribution of live data can be solved technically by an initial transfer of a copy and later regular transfer of changed data, or by giving the user access to the data the provider maintains through a network connection (so called "live data").

Distribution of data and updates is an optimization issue: how much delay between changes by the provider can the user tolerate, how often do data elements change, and how large are the meaningful data, i..e. granularity? Network access to data is technically not difficult, and the required connection to the internet is today usually in place; network access however requires very careful authorization control.

6.5 Transaction concept

Transaction management controls the effects of actions of different users and how they could interfere with each other [Gray and Reuter, 1993]. The management of transactions is necessary whenever data is shared, but often very simple solutions are sufficient, although they can sometimes be dangerous!

In case of distributions of *snapshots*: the copy that is to be distributed must be made when no changing operations are in progress.

More care is required when *live data* is distributed: changes and access must be done via a transaction management system to avoid the distribution of inconsistent data; a user who reads data must access data in a consistent state, i.e. effects of a change started after the first read must be 'held back'.

The most demanding controls are necessary when users *update* the data. Traditional databases allow updates in a transaction only if the user is connected to the database server; changes which are in conflict are thus detected and properly synchronized (i.e. forced to execute one after the other). In many GIS applications, data collection in the field updates the data base, but the user collecting the data is not connected to the server; in such cases, a novel form of transaction concept called "eventually consistent" can be applied [Vogels, 2009]. It allows updates which are not synchronized, and integration and the detection of conflicts follow later; it is possible that conflicts between teams that have collected data independently are discovered later and must be reconciled.

With an attitude of "transaction concepts are not required our application is so simple", data may be lost: the trivial transaction concept of "last wins" applies by default, and uncoordinated, later transactions wipe out previously entered data: data loss occurs!

7 DIFFERENT FORMS OF SHARING DATA

Different types of sharing data must be differentiated, to set the ground for different solutions and their applicability. Two decisions are dominant:

- is a complete data set communicated, or are smaller pieces accessed on demand?
- is changed data flowing back from user or not (two-way or one-way flow)?

Many of the differences between technical solutions for sharing data have to do with the granularity of the data: what is the unit of data which is typically interpreted, used, and transmitted? Examples for large granularity are satelite images; for small granularity, an example would be administrative applications: the name, address, and phone number of a person is a small amount of data.

The cost and delays involved in the transfer compared to the frequency of change and the difficulties of updating the data determine what the optimal solutions are. Optimality depends on the technical solutions of the time; changes in technology and new technological solutions change what arrangement is optimal.

7.1 Sharing data as backdrop and without feedback

The producer gives data or access to data to the user, but no updates flow back from the user. In general, the user does not change the data, because she may receive updated new versions without the changes she made. The arrangement is a simple “one-way street”: data flows from the producer to the user only.

7.1.1 Example: Distributing Geodata, e.g. ortophoto or images of maps, as backdrop

Many applications use satellite images, aerial photos, orthophotos or images of maps as background for their presentation of data of interest. The use of web services like Google Maps or Open Map Server is built into many web applications to serve as background to a geographic context. The spatial location is encoded as coordinate values and scale.

Semantics: The distributed data requires human interpretation and is not used for anything more than providing the context for some other data; humans are surprisingly flexible in the interpretation of images and are not confused by systematic changes of colors in images during the seasons etc.

The geographic data added to the background is registered by coordinate values; this works only if the coordinate systems are compatible and transformations between the coordinates used by the data provider and the coordinates used by the user are known.

Authorization: Data from Google Maps or OpenStreetMap [http://wiki.openstreetmap.org/wiki/Main_Page] is widely available; the restrictions are legal contracts (for open street map: http://wiki.openstreetmap.org/wiki/Open_Database_License) limiting the use.

Additionally, some checks to spot unusual behaviour and to restrict access to the web servers so that other users are not disturbed are implemented. Google may use digital watermarks to be able to trace and prove origin if its data appears somewhere without reference to the source.If images travel or are stored on media that are not well secured, they can be encrypted with secure keys, often done when the images are compressed.

Access: In general, satellite images and areal photos are very large data sets (terrabytes) and distribution is often done best by copying the data onto a medium (hard disk), which may still take several hours, and then transporting the medium physically.

7.2 Using structured geodata distributed as a snapshot and adding own content

Data about buildings with street addresses and geocodes can be used to localize other data, e.g. the location of clients of a company, or the members of a political party. Maps showing the density of clients may allow the planning of campaings, checking the service area etc.

7.2.1 Example: Street map to localize clients

If the map of clients needs to be maintained for extended periods, new clients are added and others are dropped by the user of the shared data; but the street map also changes, albeit more slowly. The changes in the street map must be introduced; in particular, newly constructed streets must be added in order to map clients from these new streets.

If the producer distributes a new snapshot from time to time the clients must be connected to the new snapshot. The identifiers used by the producer should remain the same from snapshot to snapshot to assure that the data linked to the map, i.e. linked using the identifiers, continue to work, and the user must not “relink” all the data, only the data linked to new units.

Semantics: Between the provider of the data and the user there must be a common understanding of the definition of the reference objects (e.g. streets and buildings, parcels). When adding content to data, a relationship between the added data and the data received from a producer must be established and the identifiers used (e.g. streetname and civic number) must be kept stable

Authorization: Data from public registries is often confidential (e.g. land registry) and access must be controlled; in simple cases, authorization is granted to organizations or to persons in authorized

organizations, but some data may require higher levels of control and so all access to the data must be recorded.

Access: The data can be transferred by being stored on a medium and transported in a secured manner, or it can be encrypted and transmitted over the internet. The format of the data must be communicated, and access methods prepared to read the data.

7.3 Distribution of "live data"

The producer constantly updates the data to reflect the current situation; the user always accesses the current state of the data.

7.3.1 Example: traffic data used in dispatching emergency services

Traffic data is changing rapidly, and application to dispatch service vehicles need access to current, up-to-date data. But access to current data applies equally to other, slower-changing data; examples are ownership and occupancy data in planning.

The same issues as for data distributed as snapshots apply, but there are some additional ones, mostly regarding the methods to access the data. Access to live data requires that the provider opens an access port for the user on their system; if the user is only allowed to read but not update the data, the software servicing the port must be constructed so as not to allow updates (beware of the famous SQL injection attack! [Boyd and Keromytis, 2004]). The provider will, secondly, ensure that only authorized organizations or authorized persons are granted access this is best achieved via VPN (virtual private network [http://en.wikipedia.org/wiki/Vpn]) or SSH [http://en.wikipedia.org/wiki/SSH_Communications])

7.4 Sharing with updates by the user

The user is permitted, sometimes even required, to change the data if differences to the "real" situation are discovered: either to correct errors which were present in the data previously, or because "reality" has changed. Two very similar situations occur, which are mostly differentiated by the granularity of the units of data that are subject to update. Updates of land use of individual parcels is an example with small granules and asks for conventional database transaction management; maintaining a map archive is a situation where complete larger units of data are updated.

7.4.1 Shared maitenance of structured geodata

If multiple organizations cooperate not only in the use but also in the maintenance of the shared geodata, increased efficiency is possible. Take as an example a case where the maintenance of land use data is distributed among different agencies, which are notified in certain cases: the building permit department gets informed through buidling permit applications if land use changes from agricultural to building; the agriculture department is informed by applications for subsidizes of changes from e.g. pasture to wheat growing; and the forest department collects information about logging. Together, they can maintain the land use data better.

Semantics: The classification and encoding must be integrated and agreed upon. Different departments will desire finer classifications for the parts they are interested in the joint classification must be the finest of all[Frank et al., 1997].

Authorization: records of which authorized persons caused which changes are highly recommended to avoid problems later on, when a change is questionable and responsibilities and justifications need to be found.

Access: A transaction system is necessary.

7.4.2 Shared map archive

A common situation are organizations which have a shared map archive: geodata is stored in form of plots (i.e. CAD files) and these are the units of data which are managed. In these cases, the semantics of the symbolization are typically well-standardized and the authorization rules are administratively fixed. What is missing is often a transaction management system; the rules that physical map originals automatically enforce, namely only one person can have it at a time to make changes, is removed in a digital archive. Many people can have equally "original" datasets for update, but changes applied in parallel do not get merged at

the end; only the changes of the last person to check back the updated version survive, and all others are lost – the “last wins” transaction management rule applies by default.

8 SUMMARY

Sharing geographic data requires attention, but is generally beneficial. It reduces cost and can improve the quality of the geographic data used for planning. The concerns can be differentiated in three sets of issues:

Semantics: Definitions for classifications should be established as cleanly as possible and properly documented. The approach of RDF [Manola et al., 2004] to identify different definitions with the qualification of a code seems to be more promising than the standard approach of ontologists to pretend that there is one “correct” definition. Encoding classifications by codes qualified by the definition document (and its date) makes it clear if two data sets are using different codes (perhaps only slightly different, but different still).

Authorization: Authorization documents spell out what data can be used by which persons from which organizations for which purposes. If the data is sensitive, then records of who accessed or changed data must be kept.

To ensure that authorization rules are observed, data must be encoded when traveling over the internet (VPN or SSH are good tools for this) and access control mechanism must be in place when the data is stored on machines accessible by many.

Access: Standardization of data structures is advanced and only few methods remain (e.g. for storage and compression of image data); unfortunately, some are proprietary and restricted to (expensive) software. Often access to individual parts of the data collection is possible over the web through web interfaces for relational databases or SPARQL endpoints for RDF data [Prud’Hommeaux et al., 2008].

Great attention should be given to transaction management for spatial data. The CAP (or Brewer’s) theorem[Brewer, 2012] dictates that no perfect solution fulfilling all requirements is possible: consistency at all times is only achievable if updates are only permitted if all data collections are accessible; or, update of distributed and not always connected collections is only possible if we accept a system which will “eventually” be consistend, but tolerates intermediate, non-consistent states.

9 REFERENCES

STEPHEN W BOYD AND ANGELOS D KEROMYTIS. SQL rand: Preventing SQL injection attacks. In Applied Cryptography and Network Security, pages 292–302. Springer, 2004.

ERIC BREWER. Cap twelve years later: How the" rules" have changed. Computer, 45(2):23–29, 2012.

ALEXIS COMBER, PETER FISHER, AND RICHARD WADSWORTH. Integrating land-cover data with different ontologies: identifying change from inconsistency. International Journal of Geographical Information Science, 18(7):691–708, 2004.

L. FLECK. Genesis and development of a scientific fact. University of Chicago Press, 1981.

A. U. FRANK, G. S. VOLTA, AND M. MCGRANAGHAN. Formalization of families of categorical coverages. International Journal of Geographical Information Science, 11(3):215–231, 1997.

J. GRAY AND A. REUTER. Transaction Processing: Concepts and Techniques. Morgan Kaufmann, 1993.

FRANK MANOLA, ERIC MILLER, AND BRIAN MCBRIDE. RDF primer. W3C recommendation, 10:1–107, 2004.

RYUTAROU OHBUCHI, HIROO UEDA, AND SHUH ENDOH. Robust watermarking of vector digital maps. Proceedings IEEE International Conference on Multimedia, 2002. ICME’02. volume 1, pages 577–580. IEEE, 2002.

ERIC PRUD’HOMMEAUX, ANDY SEABORNE, et al. SPARQL query language for RDF, W3C recommendation, 15, 2008.

WERNER VOGELS. Eventually consistent. Communications of the ACM, 52(1):40–44, 2009.

STUART WEIBEL, JOHN KUNZE, CARL LAGOZE, AND MISHA WOLF. Dublin core metadata for resource discovery. Internet Engineering Task Force RFC, 2413(222):132,1998.