20th International Cartographic Conference, Beijing, China, August 2001, Vol 4, pp 2804-2809

INCREMENTAL UPDATING AND VERSIONING

Paper number: 19007

Antony Cooper
icomtek, CSIR, PO Box 395 Pretoria, 0001, South Africa
Facsimile: +27 12 841 4720
acooper@csir.co.za

and

Ammatzia Peled
Department of Geography, University of Haifa, Haifa, 31905, Israel
Facsimile: +972 4 824 6814
a.peled@uvm.haifa.ac.il

Abstract

A user of geographical data invariably compiles their data base using core data sets obtained from various sources. They then integrate these together to meet their particular needs and build their own, value-added data and topology on top. The user's concern is their value-added data and maintaining their integrity, quality and spatial referencing. However, the core data provide them with a crucial spatial framework for their value-added data. As geographical data are dynamic, most suppliers of core data sets maintain and update their data. Such updates could be continuous (eg: telemetry), periodic (eg: daily, monthly or annually), part of a planned update cycle (eg: for a national mapping series), when the amount of changes crosses a threshold, or by special request (eg: census or election).

Historically, these updates have been provided to the end user in bulk, as a new data set to replace the old one. The user could ignore the update (if it is not significant enough), use it to manually (and selectively) update their data base, or accept the whole update *in toto* - and have to re-integrate it with their other core data sets and rebuild their value-added data and topology. They would also have to understand how any changes, aggregations, sub-divisions, additions or deletions might affect the geocoding or referencing of their value-added data, such as through the loss or change of an unique identifier.

The user would also need to track all the different versions they have received and used, to ensure that they implement the updates in the correct order, that they do not re-implement updates or that they do not miss updates. Unlike software, a user might need to keep several different versions of the same data set and even use them together simultaneously, such as for time series analysis.

This presentation will outline this problem of incremental updating (providing updates successively to users) and versioning (keeping track of different versions of a data set), and will attempt to provide a conceptual framework of the issues, such as:

- ❑ The data set's temporal domain – the currency or validity of the data, which can be dependent on the scale of the data;
- ❑ The cartographer's temporal domain – the period of validity of the cartographer's knowledge and understanding of the data;
- ❑ Using knowledge of the data set's and cartographer's temporal domains to resolve disputes and for historical research;
- ❑ Long transactions and parallel updates;
- ❑ When a change should be registered; and
- ❑ How incremental updating and versioning benefits the data producer.

1. What do we mean by incremental updating and versioning?

A user of geographical data invariably compiles their data base using core data sets obtained from various sources. They then integrate these together to meet their particular needs and build their own, value-added data and topology on top. The user's concern is their value-added data and maintaining their integrity, quality and spatial referencing. However, the core data provide them with a crucial spatial framework for their value-added data. As geographical data are dynamic, most suppliers of core data sets maintain and update their data.

In the analogue era, hard-copy maps themselves served as data bases for spatial data - updating the map also meant updating the spatial data set. Users would draw (or plot) their value-added data on the paper map, but the lifetime of the paper map would invariably be less than the update cycle for the map, so transferring their value-added data to the new, improved, paper map would not irk the user significantly, as they would be doing it anyway when the paper map reached the end of its useful life.

In the digital era, maps should be updated by updating the spatial data base and then producing the new maps. Updating can take place continuously (eg: via telemetry from remote sensors), periodically (eg: daily, monthly, annually, etc), as part of a planned updating cycle (with each subset of the data base or each map sheet being updated in rotation), when the amount of data to be updated and/or errors to be corrected crosses a threshold, by special request for a specific need (such as an election or census), or other reasons.

There is rapid change taking place all over the world, especially to the types of physical features recorded in spatial data bases. In addition, more and more data are captured for these and other features, as the number of spatial data users, applications and sensors grow. Thus, there is a need to speed up the frequency of updates and to automate and track the update processes. To facilitate this, one needs a formalised, continuous and incremental updating process for digital spatial data bases, and some method of keeping a record of the different versions of individual data sets, features and/or attributes. Typically, the updated versions of a data set are currently disseminated to end users by sending them the whole, updated version of the data set, or an updated subset (eg: a tile or layer). Examples of this process are NIMA's Digital Chart of the World (DCW) and road atlases for in-car navigation.

End users tend to use data sets obtained from several different sources, which they often have to integrate themselves, and upon which they build their own value-added data sets. Ultimately, the end user is more concerned about maintaining the integrity, quality and spatial-referencing of their value-added data and topology, in which they have invested much time and money, rather than the external data sets they use. Yet, these external data sets provide a crucial framework for their value-added data sets.

Hence, when a user receives one of these bulk updates today, they are faced with the dilemma of either ignoring the update (if it is not significant enough for them), manually (and maybe selectively) updating their data base based on the update they have received, or accepting the update *in toto*, but with the need to then rebuild their value-added data on top of the updated data set. This rebuilding process involves checking to see if any of the base features they have used for geocoding their non-spatial data (or other purposes) have been changed, aggregated, sub-divided or deleted, and then making the appropriate changes (which can be complex). The update might also result in the loss or change of unique identifiers, and might also require the user to rebuild the topology of their spatial data base. They also have to track which updates they have received and which they have used (and how), to ensure that they don't re-implement an update they have already implemented, miss out on crucial updates or implement updates in the wrong order.

It must be borne in mind that these problems of updating one's data set with base data received from other sources is not unique to the end users, but also applies to the producers of base data sets. Within their organisations, producers will have several field teams, photogrammetrists and other professionals updating the base data asynchronously in parallel, and hence any technologies for managing and automating the process of incremental updating and versioning will benefit producers as well.

Some work has been done on updating incrementally events, though this is done only to base data sets where the users do not add any significant amounts of value-added data. A successful example of this is the process of updating Electronic Nautical Charts (ENC), done by various national hydrographic offices around the world, under the guidance of the International Hydrographic Organisation (IHO).

It is clear that the fundamental, underlying issues of incremental updating and versioning are not unique to geographical information – other examples include solid geometry modelling and source code management systems, each with their own peculiarities [Hawla 2000].

2. The ICA Working Group on Incremental Updating and Versioning

Against this background, the International Cartographic Association (ICA) established a Working Group on Incremental Updating and Versioning at its General Assembly in Ottawa, Canada, in August 1999. The Working Group first met there in Ottawa to plan its work, and subsequently at the XIXth Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS) in Amsterdam, the Netherlands, in July 2000, held a joint ICA and ISPRS Workshop on Incremental Updating and Versioning of Spatial Data Bases. This presentation draws on these two meetings of the Working Group, especially the keynote address at the workshop [Cooper & Peled 2000].

The following are a selection of the terms of reference for the Working Group:

1. To serve as a focal point for research into the incremental updating and versioning of digital spatial data bases and the implementation of solutions. Research issues include: bi-directional, multi-level, historical and temporal updating, planning for future changes, data base maintenance, feature identifiers, modularity (dimension, context, layer, theme and size), inconsistent updating and simultaneous updating by field teams;

2. To conduct a literature study and publish an overview of the current state of the art of the incremental updating and versioning of digital spatial data bases, especially for protecting the integrity and spatial referencing of value-added data and topology;

3. To organise seminars and/or workshops in conjunction with International Cartographic Conferences and other events, leading to:
   (a) A conceptual model of incremental updating and versioning of digital spatial data bases; and
   (b) In the longer term, the publication of a reference manual (cookbook) focussing on the definition of the problem, the setting of further research goals and efforts, and identifying best practices;

4. To define algorithms for modular and/or application-oriented incremental updating and versioning;

3. Incremental updating of spatial data sets

There is probably a continuum of different ways of updating a spatial data set, from an once-off base data set that is not maintained and hence, for which there are no updates (eg: for a special event or project, or where there are no funds for maintenance), to a base data set that gets updated automatically, continuously and transparently to the user and their value-added data (utopia!). It would probably be useful to consider this continuum and draft a taxonomy or classification of the different methods of updating a spatial data set. Then, for each class we could consider what is necessary to make that method feasible, taking into account issues such as frequency of update, data volumes, security, authentication, integrity of value-added data sets, alerting users to updates, delivery mechanisms, degree of automation, metadata and other required technologies.

Further, it might be apposite to compile a glossary of the relevant terminology (perhaps in several languages), providing definitions for these terms, drawing on the work of the project team for ISO 19104, Terminology, and others. Key terms to define might include: base data set, updating, incremental updating, versioning and value-added data sets.

A good example of the problem of incremental updating is the tale of the disappearing frogs in the frorests: two forests near each other each have their own population of frogs. With time, trees grow between the two and the forests merge into one. Acknowledging this, the data supplier allocates a new identifier for the new forest, deleting the old identifiers – and in the user's data base, the frogs disappear because they no longer have a geographical reference [Bobrich 2000]!

Revisions should be done to specification, so that one can manage the process, with only the differences in data bases being updated (as opposed to a complete, undifferentiated revision of an entire map sheet, for example). The allocation of identifiers must be strictly controlled, to ensure the integrity of the data and to be able to advise the end user of exactly how the data have been changed, to facilitate the user automating the updating process as automatically as possible [Højholdt & Holme 2000]. It might also be useful to segment the data set first.

One current technique for addressing the problem of incremental updating in object-oriented data bases is long transactions, with updates occuring in parallel, which are then merged back into the main stream. One often has to resolve conflicts between the different updates, which invariably would have to be done interactively by an expert (unless one could categorically prioritise the updates *in toto*). In the process, one would maintain a version tree of all the updates, which would allow one to roll back any changes – given the volumes of data involved, a key question is how far back in the version tree should one archive the updates [Hardy & Woodsford 2000].

Whatever techniques are used, it is crucial to record and maintain the metadata for the updated data set, so that the producer can keep the user fully informed about the updates.

A perhaps contentious issue to consider would be the extent to which users are currently “locked in” to the base data sets of a specific data producer, and whether or not incremental updating and versioning will increase the user’s dependence, or provide them with independence! Could one develop common standards for incremental updating and versioning?

4. Versioning of spatial data sets

A spatial data set is a model of the real world as it is, was or might be. Two versions of a data set might differ in time, with the second version showing a more up-to-date version of the first one (we shall refer to this as being "vertical"), or they might differ in the way they are extracted from the base data set – eg: cadastral data vs transport networks (we shall refer to these differences as being "lateral").

The temporal domain, or currency, of a data set might be valid for only a brief period - a snapshot of a rapidly changing data set, such as military manoeuvres. Or, the data set’s temporal domain might remain valid for centuries or longer, such as with geological data. To some extent, the temporal domain is dependent on the scale of the data, as detail is more likely to change more quickly than generalised data: compare a town on a small-scale map, where it would be represented by a single symbol that would only change when the town’s status changed, with a cadastral map of that town, which would change every working day. We shall call this the data set’s temporal domain (DSTD for short).

However, it is sometimes overlooked that there is, effectively, a second temporal domain for each spatial data set - namely that of the cartographer or compiler of the data set. We shall call this the cartographer’s temporal domain (CTD). This second temporal domain represents that period during which the data as recorded in the data set constitute the cartographer’s knowledge and understanding of what was being modelled. Subsequently, the cartographer might discover errors in the data or have access to other source material relevant for the period being mapped, and an update for the data set might be produced. This update would then have the same DSTD as the original data set, but a more recent CTD. Naturally, when we refer to "cartographer" here, in most circumstances this would actually be a group of people - cartographers, field teams, etc, but the concept is still valid.

In general, most updates to a data set would have both more recent DSTDs and more recent CTDs than the original. An example of when an update could have a more recent DSTD but the same CTD, would be when the cartographer compiles both the original data set and the update from the same source material and with the same knowledge and understanding of what was being modelled, such as when a series of time-slices of the data are being produced together. It is quite possible that most people would think that an update represents a change to the DSTD of the data set, and only that. Probably very few people are aware of changes to the CTD of a data set or the distinction between DSTDs and CTDs.

Clearly, most users are interested in the latest available DSTD for a data set, as that represents the latest available data. However, there are many users that need older DSTDs and that need to know the CTD for a data set or update. For example, if a dispute arises over ownership of land or of the siting of a boundary, it might be necessary to know for legal reasons to resolve the dispute, what the common understanding was at the time of the transaction or when the boundary was first drawn, or indeed, at other times. A meticulous recording of all the DSTDs and CTDs might be necessary to resolve the issue to the satisfaction of all parties. CTDs are also necessary for historical research to enable understanding of what limited or incorrect model of the real world was used for making the decisions that were made at the time - an example might be understanding whether or not Christopher Columbus knew what he had found when he discovered the Americas.

To summarise, a data set is current for a particular time period and it reflects an understanding of that time period at some arbitrary moment.

Tracking different versions of a spatial data set is different from tracking different versions of a software package - invariably, once one has installed successfully a new version of a software package, one would discard the older version. There are circumstances where one might want to install, keep and track different versions of a software package on one’s computer (eg: the production and developmental versions of open source software, such as Linux), but it is unlikely that one would keep more than two or three versions, or try to use more than one version simultaneously.

However, there are many users of spatial data sets that need to keep many different versions of the same data set, such as for legal reasons, historical analysis or time-series analysis - indeed, for the latter one would also want to use several different versions of the same data set simultaneously. Clearly, we need a taxonomy or classification of the different ways of versioning spatial data sets.

5. Conclusions

There is much research to be done on incremental updating and versioning before we will see the technology embedded in systems and processes. The following is a brief summary of issues related to the incremental updating and versioning of spatial data bases that we feel need to be addressed:

- ❑ There are many spatial data sets supplied by many producers that provide users with the framework for their spatial data bases (ie: base data sets) and upon which they build their value-added data sets and topology;

- ❑ There should be no need to redistribute an entire data set to its users to propagate changes that are only minor or few in number;

- ❑ It will be more efficient to disseminate only the changed or updated data (ie: patches to the data set);

- ❑ There is a need for secure algorithms and procedures for updating incrementally a base data set, to minimise the impact of the changes on the users' value-added data sets and topology;

- ❑ It will be more efficient and effective to automate the update process, to avoid error-prone, labour-intensive work to introduce the changes to one's existing data bases;

- ❑ Data producers run out of budget, resulting in inconsistent updating of base data sets, not according to plan;

- ❑ There is a need to keep track of different versions of a data set, for legal reasons, time-series analysis, historical research and planning;

- ❑ There is a need for identifying properly different versions of a data set, to ensure that updates are implemented properly and in the correct sequence;

- ❑ Generalization will become the way of producing special-purpose products from a base data set, and not just a means of producing the data at a smaller scale; and

- ❑ There is a need to define carefully the concepts of incremental updating and versioning – we do not yet have a common understanding of them!

To finish with a question from the Working Group's workshop in July 2000: does an update to a data set or a new version necessarily mean an improvement to the data set?

6. Acknowledgements

We would like to acknowledge the contribution that the ICA has made in establishing the Working Group on Incremental Updating and Versioning, and we would like to acknowledge the contributions made by the participants in the meetings of the Working Group. We would also like to acknowledge the financial support of our institutions, which made this presentation possible.

7. References

[Bobrich 2000] **Bobrich J**, July 2000, *Fusion and Incremental Updating of Cartographic Generalized Data*, Joint ICA and ISPRS Workshop on Incremental Updating and Versioning of Spatial Data Bases, Amsterdam, the Netherlands.

[Cooper & Peled 2000] **Cooper AK & Peled A**, July 2000, *The International Cartographic Association's Working Group on Incremental Updating and Versioning*, Joint ICA and ISPRS Workshop on Incremental Updating and Versioning of Spatial Data Bases, Amsterdam, the Netherlands.

[Hardy & Woodsford 2000] **Hardy P & Woordsford P**, July 2000, *Incremental Updating Using the GOTHIC Versioned Object Database with the Hydrographic S57 ENC and SOTF Spatial Object Transfer Formats*, Joint ICA and ISPRS Workshop on Incremental Updating and Versioning of Spatial Data Bases, Amsterdam, the Netherlands.

[Hawla 2000] **Hawla D**, July 2000, *pers comm.*

[Højholdt & Holme 2000] **Højholdt P & Holme D**, July 2000, *Revision of Maps Registrating Only True Changes*, Joint ICA and ISPRS Workshop on Incremental Updating and Versioning of Spatial Data Bases, Amsterdam, the Netherlands.