Merging Different Data Sets Based on Matching and Adjustment Techniques

Lothar GRUENDIG, Frank GIELSDORF, **Bernd ASCHOFF**, **Germany**

Key words: digital cadastre, geo information, data integration

SUMMARY

The paper presents a general approach for a matching based on geometrical and topological information. A solution for the extraction of sub graphs based on adjacency tensors is presented. Based on such sub graphs matching operators with different levels of complexity can be defined. The search for matching candidates is supported by very efficient algorithms based on n-dimensional search trees. Detected matching candidates are checked via statistical tests, and thereafter remaining ambiguities are removed.

The results of the matching process are observations, for instance point identities, which are introduced in the adjustment process of merging to follow.

Presented examples show that matching and adjustment can not be seen as independent work steps, both tasks are part of one superordinate iteration process.

ZUSAMMENFASSUNG

Im vorliegenden Beitrag wird ein allgemeiner Matching-Ansatz, beruhend auf topologischen und geometrischen Informationen, vorgestellt. Dies beinhaltet ein Verfahren zur Extraktion von Teilgraphen mit Hilfe von Adjazenztensoren. Auf der Grundlage solcher Teilgraphen können Matching-Operatoren von unterschiedlicher Komplexität definiert werden. Die Suche nach Matching-Kandidaten wird durch effiziente Algorithmen, aufbauend auf n-dimensionalen Suchbäumen, unterstützt. Die gefundenen Matching-Kandidaten werden mittels statistischer Tests überprüft und verbleibende Mehrdeutigkeiten werden entfernt.
Das Ergebnis des Matching-Prozesses sind Beobachtungen, z.B. Punktidentitäten, welche in ein Ausgleichungsmodell eingeführt werden.

Anhand von Beispielen wird gezeigt, dass Matching und Ausgleichung nicht als unabhängige Arbeitsschritte betrachtet werden können. Vielmehr sind beide Schritte Bestandteil eines übergeordneten Iterationsprozesses.

Merging Different Data Sets Based on Matching and Adjustment Techniques

Lothar GRUENDIG, Frank GIELSDORF, Bernd ASCHOFF, Germany

1 INTRODUCTION

A necessary requirement for GIS use and maintenance is to be able to integrate data sets of different quality and origin. This is the case when data with high geometric accuracy, for instance GPS points, are integrated in data sets which trace back to digitization of analogue maps. But also merging of data from different web map servers can lead to massive problems because of the inhomogeneity of the data involved.
These problems can be solved by applying an adjustment process for merging which models the accuracy and the distance dependent correlations of the coordinates. However, a prerequisite for such an adjustment process is the availability of information about identical features, in particularly points. This information is commonly not explicitly stored. The only way to get this information is therefore to apply a matching process based on the properties of the map features. Frequently the semantics of the datasets is different, so that only geometrical and topological properties remain as common characteristics for the matching process.

2 ADJUSTMENT

2.1 Modeling of Correlations

In the very most cases the geometry of spatial vector data is parameterized by point coordinates. These coordinates might have descended from different sources. Frequently they are the result of a digitalization process of the basis of analogue paper maps. But also orthophotos, classical surveying or GNSS can be the basis of the data. The technology applied for the coordinate production determines their geometrical accuracy.

In the past the field surveying as well as the plotting of the maps where done following the principle of neighborhood. The application of this principle leads to significant distance dependant correlations between the coordinates which are therefore present in many spatial datasets.

During the integration of two spatial data sets these correlation have to be taken into account. A typical example is the introduction of very accurate GPS coordinates in a data set derived from analogue maps.
Figure 2-1 shows the effect of neglecting the distance dependant correlations between coordinates.

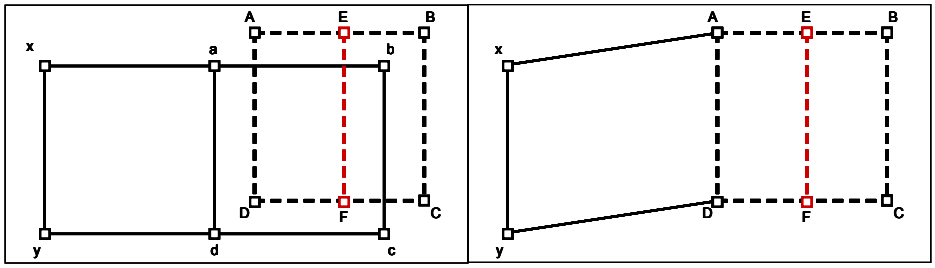

Figure 2-1: (Left) cadastral boundaries (solid lines) and higher accuracy sub-division (dashed lines). (Right) polygon xady is distorted if points abcd are replaced by ABCD.

In this example the GPS points A, B, C and D are introduced in a map based cadastral data set. Neglecting the correlations leads to a massive distortion of the parcel XADY.

In general the exact variances and covariances of a coordinate set are unknown. Commonly one has just very general information about the point accuracy. Distance dependent correlations can only be modeled on the basis of hypothesis. There exist two general ways to model distance dependant correlations: The stochastic and the functional approach.

The stochastic approach uses a correlation function to model the correlation between two point coordinates as function of the point distance. Commonly used models are:

$$\rho_{ij} = \frac{1}{1+\frac{d_{ij}}{d_0}} \quad \text{or} \quad \rho_{ij} = \frac{1}{1+\left(\frac{d_{ij}}{d_0}\right)^2} \tag{2.1}$$

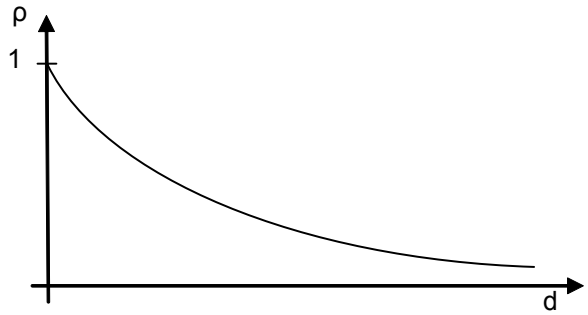

Figure 2-2: Graph of a distance dependant correlation function

The disadvantage of this approach is the fact that it can lead to very large covariance matrices. Therefore often a maximal distance is set for the purpose of limiting the area of covariances to be applied.

In the functional approach the correlations are modeled by artificial pseudo observations [1]. The pseudo observations are coordinate differences between neighboring points. The neighboring information can be derived from a Delauney triangulation. The weight of the pseudo observations can be chosen in such a way that the behavior of a rubber membrane will be simulated.

Both approaches cause a conformal behavior of the adjacent region if a point is shifted.

2.2 Identities

A prerequisite for any integration of different spatial vector data sets is the information about identical objects in these data sets. The most common and even the simplest case is to model the identity of points.
Often such an identity is modeled by one point object carrying two or more sets of coordinates. This way is easy to realize in an adjustment program but it has essential disadvantages. During the matching process in which identity information is generated misidentifications frequently occur. If the identity information is modeled as described above then it becomes very difficult to detect a false identification in the adjustment result. Even if it could be detected it would be later on necessary to split the detected point object which could lead to severe data management problems.

A more sophisticated way is the modeling of point identities by identity observation. An identity observation contains nothing else but two coordinate differences with an observation value of 0. The point identities can be weighted similar to the coordinates itself. During an adjustment based analysis misidentifications can be detected easily. They can be corrected by deleting the relevant identity observation. After the adjustment process the remaining identical points are fused to unique objects.

2.3 Geometrical Constraints

Frequently vector data sets imply several geometric constraints. Typical examples are rectangularities for buildings or collinearities for border lines. If information about such constraints is available then it can be considered in the adjustment model.

One option is the formulation of constraint equations in the adjustment model. But this solution doesn't allow the detection of possibly wrong constraints. More practicable is the modeling of constraints by observations. Like ordinary observations they can undergo a blunder detection process by means of normalized residuals.

In program system SYSTRA of the authors rectangularity is modeled by scalar product observations, collinearity by cross product observations. The weight of these observations is derived from coordinate accuracies applying variance propagation which means that the resulting weights are appropriate for any arbitrary point configuration.

Further types of implemented geometrical constraints are parallelities with or without defined distance, circle continuities and others.

3 MATCHING

3.2 General Approach

The prerequisite for applying adjustment techniques for the integration of spatial vector data is information about identical objekts. The process to find such identical objects in different datasets is called matching. The matching process results in identity and constraint observations which are introduced in an adjustment calculation.

Topological Properties

The information contained in spatial vector data can be subdivided in a topological and a geometrical part. The topology can be seen as a 2D graph with vertices, edges and meshes as basic elements. The following operators are based on the definition of sub-graphs. A sub-graph is a well defined combination of a fixed number of topological elements.

The simplest cases of sub-graphs are single vertices or edges. The next level of complexity is represented by corners. In a topological sense, a corner is a set of two non-identical edges with one common vertex. For example, a vertex with four adjacent edges provides six topological corners (Figure 3-1). Defining topological corners provides the option to match, for example, the corner of a house with the corner of a parcel, even when the house vertex and the parcel vertex have a different number of adjacent edges.

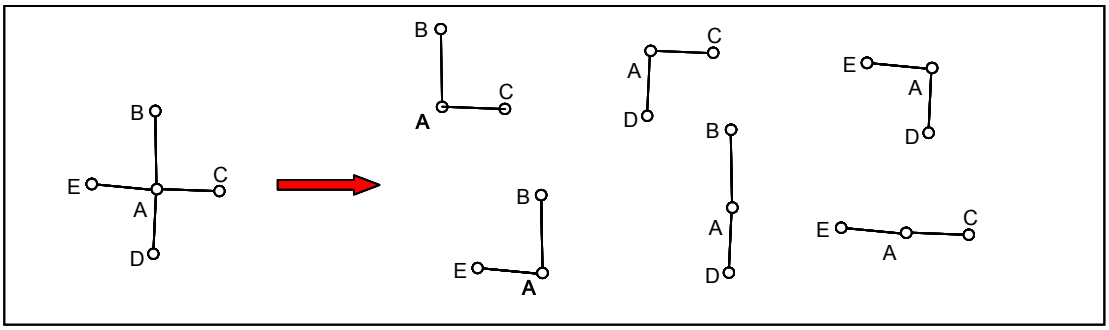

Figure 3-1: Definition of topological corners

Topological sub-graphs within a dataset can be found with a very efficient algorithm based on adjacency tensors in sparse notation.

Geometrical Properties

Each sub-graph can carry different combinations of geometrical parameters. An edge for instance can be parameterized by the coordinates of the adjacent vertices but also by the coordinates of one of the adjacent vertices and a grid bearing or by the components of its normal vector and its orthogonal distance from the coordinate origin. Each of the listed options maps a special issue.

A sub-graph parameterized by *k* geometrical parameters can be seen as point in *k*-dimensional space with each parameter as coordinate component. In the general case the parameters are correlated random values. So the geometrical properties of a sub-graph are given by its parameter vector $\mathbf{x}$ and its covariance matrix $\mathbf{C}_{xx}$ whereby $\mathbf{C}_{xx}$ can be fully occupied and singular.

Geometrical parameters can be classified by the criterion of datum dependency. Coordinates, grid bearings, normal vectors etc. are datum dependent. Angles and distance relations are datum independent (if we restrict our consideration to Euclidian spaces). If the processed data sets are given in different reference frames then identical sub-graphs can only be found with a set of datum independent parameters. In the other case of a common reference frame the search with datum dependent parameters is more efficient.

Frequently a preprocessing is necessary in which a datum independent search finds identical objects as basis for an initial datum transformation.

Finding Matching Candidates

Two sub-graphs can be accepted as identical if their distance in the belonging *k*-dimensional parameter space is small enough. But it doesn't make sense to test each sub-graph of one data set with each sub-graph of the same type in the other data set. Using such an approach the calculation effort would grow with the square of the number of the tested objects which is not acceptable in practice. Therefore it is necessary to limit the number of potential matching candidates by an efficient search algorithm.
Such an algorithm is realized using a *k*-dimensional window search supported by a *k*-dimensional search tree.

Statistical Test

Each point in the *k*-dimensional parameter space is compared with all points of the second data set falling in a predefined *k*-dimensional window. The test value is calculated from the parameter difference vector $\Delta\mathbf{x}$ and its covariance matrix $\mathbf{C}_{\Delta x \Delta x}$.

$$\begin{aligned} \Delta\mathbf{x} &= \mathbf{x}_1 - \mathbf{x}_2 \\ \mathbf{C}_{\Delta x \Delta x} &= \mathbf{C}_{xx1} + \mathbf{C}_{xx2} \\ \chi^2 &= \Delta\mathbf{x}^T \cdot \mathbf{C}_{\Delta x \Delta x}^{-1} \cdot \Delta\mathbf{x} \end{aligned} \tag{3.1}$$

The test value is χ^2-distibuted whereby the degree of freedom corresponds to the rank of $\mathbf{C}_{\Delta x \Delta x}$.

3.3 Examples of Matching Operators

3.3.1 Corner Operator

The corner operator uses corner sub-graphs. The geometrical parameterization is given by the coordinates of the corner vertex and two normalized direction vectors. This combination of geometrical parameters allows the matching of vertices even if the adjacent edges have different length. It is for instance possible to match building corners with parcel corners (Figure 3-2).

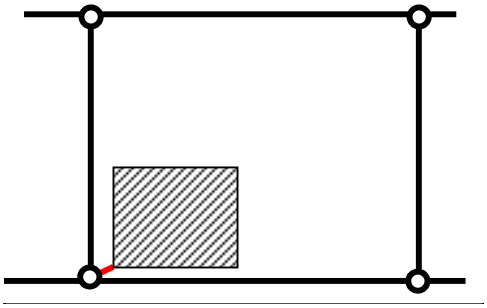

Figure 3-2: Building corner matches a parcel corner

The parameter space has in this particular case 6 dimensions (x,y,r_{x1},r_{y1},r_{x2},r_{y2}). The covariance matrix of the parameters is singular with a rank defect of 2.

3.3.2 Point Operator

The point operator uses just the coordinates of two points to compare their position. The parameter vector has the dimension 2, the covariance matrix is regular.

3.3.3 Straight Line Operator

The straight line operator uses the normal vector and the orthogonal distance from the coordinate origin to describe the geometry of a straight line. It is possible to match lines even if they have different end points. The parameter vector is 3-dimensional. The covariance matrix is singular with a rank defect of 1.
The candidate search can be limited with the restriction that the participating lines have to overlap.

3.4 Processing of Ambiguities

The algorithm implemented in the program SYSmatch follows a conservative strategy. Only if an identity is unique then the belonging observation is created. In the case of an ambiguity a warning is generated.

Experiences show that it is a practical way to begin the matching-adjustment process with operators of higher complexity as for instance the corner operator. These operators lead to a

smaller ratio of ambiguities. After some iteration steps of matching and adjustment the coordinates have an improved accuracy that allows the usage of operators with a lower level of complexity as for instance the point operator.

3.5 Finding Constraints

Different from the matching operators which compare topology and geometry between different datasets the constraint searching operators work within one single dataset. It can happen that during the integration process the shape of objects get distorted. Examples are the rectangularity of buildings or the collinearity of boundary points.
Basis for the search for these constraints are sub-graphs of the corner type within one dataset.

3.5.1 Rectangularities

The test value for the rectangularity check is the scalar product of the both corner directions. The variance of this value is calculated by variance propagation from the point coordinate variances of the participating points. The value of the scalar product is normal distributed and can easily be checked for significance.

3.5.2 Alignments

The test for collinearity of three points works similar to the test for rectangularity. The test value here is the cross product of the corner vectors.

4 CASE STUDIES

The following two projects reflect the operability and efficiency of the described data integration strategy. Connected to different database systems, a significant improvement of the point coordinates was achieved by applying the adjustment tool SYSTRA of technet GmbH. Accordingly, matching and search operators were used to generate map connections or geometrical constraints automatically in order to maintain object shapes.

4.2 Victoria, Australia

In 2006, LogicaCMG of Australia processed the “Wimmera Mallee Spatial Upgrade Study” for the Department of Sustainability and Environment (DSE) in Victoria, Australia [2]. The project objective was to improve spatial accuracy of parts of the Vitorian cadastre. The project area contained approximately 100.000 points. The dataset was sourced from a 1:25000 scale map, 90% of the cadastral points are within 25m of their true position.
Within the “Wimmera Mallee Pipeline Project” being responsible for the improvement motivation, about 14.000 GPS points of fence corners or fence alignments were measured. For the adjustment process, the point accuracy was set to a few decimeters with respect to their rectification in the cadastral map.
The Cooperative Research Centre for Spatial Information (CRC-SI) at Melbourne University and the Technical University of Berlin supported the project. The described matching and adjustment strategy was applied.

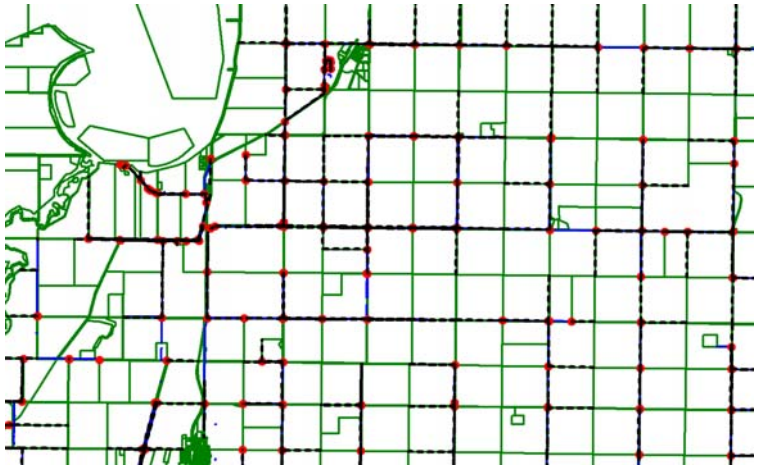

Figure 4-1: Section of the Wimmera Mallee project

Figure 4-1 shows point and fence line identities as the result of the matching process within a small section of the project area. The result of the adjustment was a point standard deviation of 1.5 m which, assuming a normal circular error distribution, relates to 90% of points being within 3.2m of their true position.

4.2 Brandenburg, Germany

After having successfully finished the project "Forced Implementation of the Real Estate Map (ALK)" in the German federal country Brandenburg, the responsible authority "Landesvermessung and Geoinformation Brandenburg (LGB)" initiated the project "Quality Improvement of the automated Real Estate Map". With respect to the limited achieved point accuracy of not better than about 1m, which resulted from the only use of analogue insular real estate maps (scales 1:2000 or less accurate), it was decided in 2006 to integrate accurate GPS observations and measurements from field book.

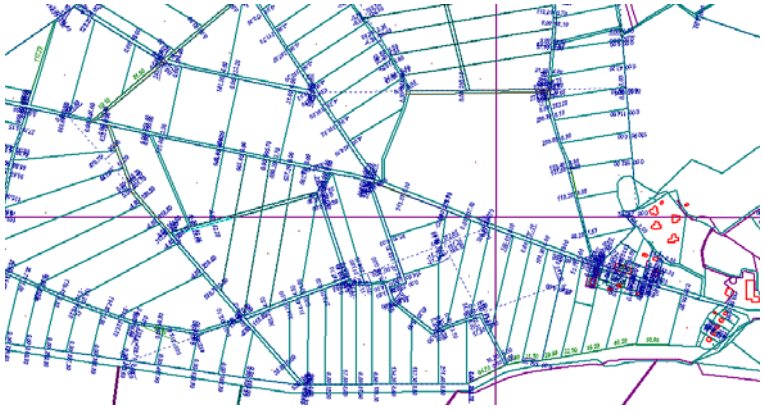

Figure 4-2: Integration of field book measures in a Brandenburg project (section)

Figure 4-2 shows the integration of measurements from field book to which a standard deviation of 2-3 cm was assigned. Based on fixed data block boundaries and a selected number of GPS measurements (0,5-1,0 cm), the resulting points coordinates would be improved by adjustment to an accuracy up to a 1-5 cm.

If points were not connected with these precise measurement data, geometrical improvement could be derived from orthophoto data of low scales. Additionally artificial constraints like rectangularity for buildings or alignment for border lines are to be generated automatically.

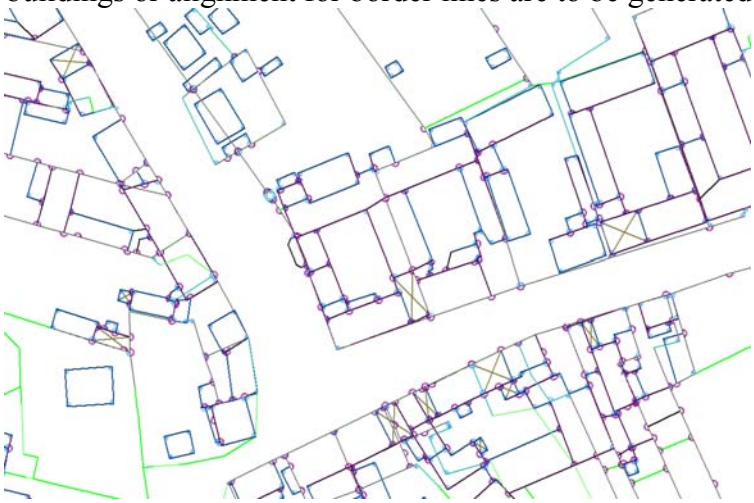

Figure 4-3: Automatically generated constraints in a Brandenburg project (section)

Figure 4-3 shows the result of the rectangularity and alignment search operator after scanning the building layer of the ALK database. The resulting measurements are able to sufficiently model or even to improve the digitized object shapes by means of the relative accuracy of the related points.

The "Quality improvement campaign" of the LGB will end in 2010 and will then be available for the whole country.

5. CONCLUSIONS

In the paper it is shown that Geo data integration is an adjustment problem. The application of adjustment techniques leads to a significant improvement of the geometrical accuracy with a relatively small effort. Only via adjustment techniques it becomes possible to integrate measurement types which already exist and are of high economic value. Adjustment models need information about identical objects. It is shown that identities can be generated automatically by efficient matching algorithms, and that sophisticated matching algorithms can be efficiently based on the theory of mathematical statistics. However, data integration can also be seen as an ongoing process where distance dependent correlations play an important role.

REFERENCES

Aschoff, B.: „Geometrical Improvement for GeoData", Workshop Quality Management for Geodata, FIG Congress 2006, Munich, 9th October 2006

Menner, G.: „Wimmera Mallee Spatial Upgrade Study“, LogicaCMG for the Department of Sustainability and Environment of Victoria, Australia, Internal paper, 6th October 2006.
Gielsdorf, F., Gruendig, L., Aschoff, B.: Positional Accuracy Improvement - A necessary tool for updating and integration of GIS data, FIG Working Week, Proceedings, Athen, 22.-27.05.2004
Gielsdorf, F.: Georeferencing of Analogue Maps via Interconnected Transformation Fourth Turkish-German Joint Geodetic Days, Proceedings, Berlin, 03.-06.04.2001
Gründig, L., Gielsdorf, F., Scheu, M., Dreesmann, R.,: ‘Umsetzung der analogen Liegenschaftskarten in die digitale Liegenschaftskarte im ländlichen Raum”, Surveying Department Brandenburg, internal paper (1999)
Gielsdorf, F.: „Nachbarschaftstreue Anpassung auf der Basis des Membranmodells“, Zeitschrift für Vermessungswesen, Heft 5 1997

BIOGRAPHICAL NOTES

Prof. Dr. Lothar Gruendig, born in 1944. Graduated in 1970 as Dipl.-Ing. in Surveying and obtaining doctorate degree in 1975, both from University of Stuttgart, until 1987 senior research assistant at University of Stuttgart. Since 1988 Professor of Geodesy and Adjustment Techniques. Head of the Department of Geodesy and Geoinformation Science, Technical University of Berlin.

Dr. Frank Gielsdorf, born 1960. Graduated in 1987 as Dipl.-Ing. in Surveying from Technical University of Dresden. Obtaining doctoral degree in 1997 from Technical University of Berlin. 1995-2006 Assistant Professor at the Department of Geodesy and Geoinformation, Technical University of Berlin. Since May 2006 director of innovative developments with technet GmbH Berlin.

Bernd Aschoff, born 1962. Graduated in 1987 as Dipl.-Ing. in Surveying from Technical University of Berlin. Since 1995 managing director of the technet GmbH.

CONTACTS

Prof. Lothar Gruendig
Technische Universität Berlin, Sekretatiat H20
Strasse des 17. Juni 135
10623 Berlin
Germany
Tel.: +49 30 31422375
Fax: +49 30 31421119
Email: gruendig@inge3.bv.tu-berlin.de
Web site: www.survey.tu-berlin.de

Dr.-Ing. Frank Gielsdorf
technet GmbH gruendig+partner
Maassenstrasse 14
10777 Berlin
Tel. + 49 30 2154020
Fax + 49 30 2154027
Email: frank.gielsdorf@technet-gmbh.com
Web: www.technet-gmbh.com

Bernd Aschoff
technet GmbH gruendig+partner
Massenstrasse 14
10777 Berlin
Germany
Tel.: + 49 30 215 4020
Fax: + 49 30 215 4027
Email: bernd.aschoff@technet-gmbh.com
Web site: www.technet-gmbh.com