Dagstuhl seminar 09161 on

Generalization of Spatial Information

14-17 April 2009

Summary Report

M. Sester, P. van Oosterom, F. van Harmelen, S. Mustière

More details on the seminar website:
http://www.dagstuhl.de/no_cache/en/program/calendar/semhp/?semnr=09161

Motivation

From the early start of handling geo-information in a digital environments, it has been attempted to automate the process of generalization of geographic information. Traditionally for the production of different map scale series, but more and more also in other contexts, such as the desktop/web /mobile use of geo-information, in order to allow to process, handle and understand possibly huge masses of data. Generalization is the process responsible for generating visualizations or geographic databases at coarser levels-of-detail than the original source database, while retaining essential characteristics of the underlying geographic information.

All current solutions for supporting different levels-of-detail are based on (static) copies that are (redundantly) stored at these levels. This makes dynamically adapting the map to new information and to the changing context of the user impossible. Besides the classic geo-information visualization requirement (supporting different scales), which has been solved only partly, there are also new requirements for generalization, making it even more difficult: it should be dynamic and suitable for progressive transfer. Furthermore, the objects to be visualized have expanded in dimension: the emerging 3D and temporal data.

In order to make further progress in automated, machine generalization both the semantics of the spatial information and the user needs should be (further) formalized. Methods and techniques from the semantic web might be useful in this formalization and tools from knowledge engineering could be applied in the actual generalization based on reasoning. Interpretation of spatial constellations or situations is a process that is closely linked to human capabilities and can be formalized using formal semantics (OWL, ODM, etc.). Making implicit information explicit is needed not only for many spatial analysis problems, but also for aspects of information communication.

Spatial data also pose exciting questions for the algorithms and data structuring communities. It is vital that computational geometers meet with the spatial data community to exchange ideas, pose problems and offer solutions. Most algorithmic problems arising in that field are indeed geometric. In this context it must be noticed that the focus is more and more on 3D (and 3D plus time) geometric computations. In this respect, generalization operations and the resulting data have to be understood as processes, which will allow a broader and more flexible usage and re-generalization when changes in reality have occurred.

For a mass market (e.g. consumers of mobile maps) the human factors aspect is very important. The currently available (often mobile maps) solutions still have insufficient user-interfaces. Extremely important is the issue of context as the user gets ‘lost’ very easily on the small mobile displays when zooming and panning. Based on a selection of use cases (navigation, tourist support, etc.), User-Centered Design techniques should be applied to evaluate the interaction and the quality of the maps.

In this context, the main goal of the seminar was to bring together experts from digital cartography, knowledge engineering, computational geometry, computer graphics and cognitive science, to lead to a fruitful exchange of different – but very close – disciplines and hopefully to the creation of new collaborations.

Attendance

The seminar brought together experts from digital cartography, knowledge engineering, computational geometry and cognitive science, with the goal to lead to a fruitful exchange of different – but very close – disciplines and to the creation of new collaborations.

The following 34 attendees did actively particpate to the seminar:
Nico Bakker, Claus Brenner, Dirk Burghardt, Omair Zubair Chaudhry, Leila De Floriani, Ahmet Ozgur Dogru, Cecile Duchene, Andrew U. Frank, Julien Gaffuri, Willem Robert van Hage, Paul Hardy, Frank van Harmelen, Lars Harrie, Jan-Henrik Haunert, Peter Hojholt, Marc van Kreveld, Werner Kuhn, Patrick Lüscher, William Mackaness, Martijn Meijers, Sebastien Mustiere, Moritz Neun, Peter van Oosterom, Nicolas Regnauld, Patrick Revell, Monika Sester, Lawrence Stanislawski, Jantien Stoter, Heiner Stuckenschmidt, Robert Weibel, Stephan Winter, Alexander Wolff, Anna Zhdanova, Esteban Zimanyi.

Presentations and discussions

Tuesday April 14

Peter van Oosterom welcomed everybody on behalf of the organizers. He outlined the focus of the seminar and pointed out the importance of the integration of members from two disciplines – where only a small group considered themselves as belonging to both communities. Therefore, the organizers had decided to have introductory talks which give insight into the two fields Generalization and Semantic Web technologies. The first session was dedicated to the introduction of all participants. Everybody had the chance to present him/herself briefly, and also indicate the respective research interest also with respect to the seminar.

The first talk of the second session was the Tutorial on Generalization Research Issues, given by **William Mackaness**. He first outlined the successes of generalization; here he listed the separation and definition of Digital Landscape Models (DLM) and Digital Cartographic Models (DCM), describing the spatial objects of the environment in a phenomenon-centric view. Furthermore, he highlighted the development of rich toolboxes with generalization operations and analysis operations. Then he moved to the challenges. Among them he pointed out the issue of duplication of research and development, which could be alleviated using common test beds, and/or shared operations via web services. An important challenge is the change in the paradigms of map use, generated by new presentation methods (3D, tactile, acoustics but also technological changes, e.g. created by the Web2.0, or new applications like on-line or ad hoc maps. Also, changes in data capture and new sensors pose new demands on generalization. These challenges lead to a re-thinking of the objectives in map generalization, going from a mere cartographic and analogue map centred view to more general paradigms like “abstraction of space in reaction to tasks” and “to reveal patterns and associations” in the spatial information, or in general “to make sense of things”.

The second talk was given by **Robert Weibel**. He presented a SWOT analysis of Map Generalization. Firstly, he presented the new situation with respect to spatial data and their presence in everyday environments. The data sources leading to massive data – often their geo-relevance is not even directly mentioned. Also, there are new methods to deal with these data. The question is, where map generalization is needed in such a new situation. He listed strengths and weaknesses, as well as threats and opportunities. Just to name some of them: one strength lies in the strong community that has tackled relevant research issues in the past, which lead to considerable advance of science and development. On the weakness side is the fact that much of the research had been driven by changes in the map production scheme of Mapping Agencies, and also related to paper maps. The threats could be that once the NMA problem – which is innate challenge of the own community – is solved, there might be a reluctance to tackle also new problems. Still, however, the challenges posed in the introduction give a lot of possibilities which should be talked with joint efforts.

Cecile Duchene gave a talk on „Use of agents in generalization". She showed the research that has been conducted in the course of the EU-project AGENT, as well as new development since then, including the integration of 2.5D-representations that have been investigated in a PhD at the IGN by Julien Gaffuri. Julien also gave a demo which showed some agents in action.

Jan Haunert talked about "Map Generalisation by Combinatorial Optimisation: Handling Multiple Constraints and Objectives". He introduced the map generalization problem as combinatorial optimization problem, where mainly two techniques are applied: deterministic methods or iterative meta-heuristics. He presented typical generalisation problems and formalised them in terms of constraints and objectives, often also termed hard and soft constraints. It was very illustrative to have this clear distinction, which also made sure that the terms in the later presentations were used accordingly. Finally, he highlighted three research challenges:
(1) We need deterministic methods that allow for a higher variability of the output maps than the existing deterministic methods do. (2) Methods based on meta-heuristic need to be improved with respect to their capability of handling hard constraints. (3) Methods for integrating multiple (potentially contradicting) generalisation objectives are needed. In the discussion, for the latter issue Multi-Criteria-Optimization was suggested as one possibility.

Alexander Wolff presented a paper on "Optimal Simplification of Building Ground Plans", which he had produced together with Jan Haunert, where combinatorial optimization was used. The approach allows simplifying building ground plans by removing and extending / intersecting edges. Also, there are rules that determine, which possible "shortcuts" from one edge to a neighboring one are possible, also taking topological aspects into account.

Nico Backer gave an overview on the generalization aspects at his institution, The Topografische Dienst / Kadaster of the Netherlands. Up to the 90s of the previous century this entirely process was manual, based on given generalization rules. These generalisation rules have been based on both national and international specifications, partly originating from military map products. The early attempts and research to automate the map generalisation started in 1990 which the idea to develop an expert system with the rules for generalisation. In the mid-nineties all the generalisation rules were analyzed and some tests were performed with special software. It was not successfully enough, therefore the decision was taken to change the base material TOP10vector to object-oriented database which should better support the automated process. In the last decade much research was done to develop automated generalisation. Until now only some functionality is available, but there are not yet implemented in the production process. The last development is to harmonize the existing data models of the current product at different scales to one information model and to convert all the vector datasets to object-oriented datasets.

The presentation of **Lars Harrie** was about "Methods to measure map readability". The creation of maps in real-time web services introduces challenges concerning map readability. Thus analytical measures are needed to control the readability. The aim of the presentation was to develop and evaluate analytical readability measures with the help of user tests.

Jantien Stoter talked about "Formalising generalisation requirements in a Multi-Scale Information Model Topography". As a prerequisite for automatic generalization, a formal specification of the content and meaning of data sets representing data in different scales is needed. In the presentation, Jantien showed modelling principles for a multi-scale Information Model TOPography, called IMTOP (studied in collaboration with TU Delft and Dutch Kadaster). The Unified Modelling Language (UML), including the Object Constraint Language (OCL), is used for formalisation. The modelling principles are a result of two steps: firstly they conducted a requirement analysis using the needs for a multi-scale information model of the Netherlands' Kadaster as case study. Secondly, they designed, implemented and evaluated several alternatives to meet the identified requirements. The model covers two main aspects of a multi-scale data model. Firstly it integrates the data states at the different scales in a UML class diagram. Secondly it formalises requirements for automated generalisation by means of OCL expressions. These requirements cover both legibility conditions regarding the output of generalisation and preservation conditions.

Paul Hardy gave a talk entitled “Optimization and Pattern Identification for Generalization”. ESRI has been developing a research prototype of an 'Optimizer' engine, focused on contextual generalization [Monnot, Hardy, Lee 2007]. It uses Simulated Annealing, but with intelligent rather than random actions, in a MCMC-fashion. He showed early example outputs from initial experimental tools using this optimization engine, for road network thinning (pruning); road displacement to avoid symbology conflicts; displacement propagation; and building displacement/exaggeration. Finally it mentions other ESRI research on detection of patterns of multiple urban buildings for enrichment prior to generalization, plus new tools (non-Optimizer) for collapse of polygon to skeletal centerline.

Moritz Neun gave a presentation where he showed the new developments of ESRI software that were developed especially with respect to the requirements of a customer SwissTopo.

In summary, the first day was mainly dedicated to one topic – generalization. Thus there were limited possibilities for contributions of the semantic web group. However, it has highlighted first of all the research challenges, and also pointed out ways ahead. Also, important paradigms were named, namely optimization techniques and constraint modelling, which have great potential to be methods that lead the research one step further.

Wednesday April 15

The Wednesday was specifically devoted to the role of semantics in geo-information.

Frank van Harmelen (Vrije Universiteit Amsterdam) opened the day with a “Bluffer’s Guide to the Semantic Web”, covering both the general principles (meta-data, URLs, ontologies), as well as specific technologies (RDF, OWL), and currently succesful applications in a number of sectors. Van Harmelen also tried to suggest how the datamodels and technologies that have proven to be succesful in other sectors (life-sciences, cultural heritage, broadcasting) could be ported to the geo-information domain. The talk aimed to do two things: disseminate basic information about modern semantic technologies to a geo-information audience, and set the scene for a discussion on the relevance and applicability of semantic technologies for geo-information.

Heiner Stuckenschmidt (Mannheim) presented a specific approach to the integration of spatial reasoning with terminological reasoning (which is a central element in current semantic technologies). This problems have been mostly mutually ignoring each other: spatio-temporal reasoning is widely neglected on the Semantic Web (existing semantic web languages are not well suited for representing these aspects), and only few geo-spatial applications exploit much semantics. Stuckenschmidt advocated an approach in which the two modes of reasoning remained separate modules that interacted in well-delineated ways. (This proposal differs from the much tighter integration found in other systems that combine spatial and terminological reasoning).

Patrick Luescher (Zurich) proposed the use of ontologies to bridge the gap between the individual, discrete objects represented in current cartographic databases and the higher order geographic concepts that humans use to reason about space, specifically in an urban environment. An ontology-based reasoner is then linked to spatial data processing capabilities through a service oriented architecture, closely in line with the observations from Stuckenschmidt’s talk. The presentation and approach were warmly applauded by the semanticists in the audience.

In contrast, **Willem ten Hage** (Amsterdam) described his work on a SWI-Prolog spatial indexing package. which allows a **tight** integration between RDF(S) and OWL reasoning and spatial reasoning (e.g. incremental nearest neighbor, containment, intersection).

Nico Bakker (Dutch cartograhpic agency) talked about the efforts athe Dutch National Topographical Service to automate map generalisation. The going has been tough, and until now only some functionalities are available, but there are not yet implemented in the production process. The presentation gave rise to a discussion wether automated generalisation was possible at all, e.g. contrasting Bakker's talk with the view proposed by Brenner in his talk on "meaningless maps".

Ahmet Ozgur Dogru presented work to support the classification of road interchanges by capturing them in a formal representation. The classifcation and underlying formal representation captured the ***topological*** nature of interchangs (using matrix and tree data-structures). This proved a useful basis for generalising such interchanges into a limited number of types each with a simplified visual representation. No further ***semantic*** interpretation was given to the geospatial features.

Andrew Frank (Vienna) gave a stimulating formal account of the fundamental process of generalisation, in an attempt to uncover why the map-generalisation is so hard, and why it has escaped a generic solution for over 40 years now. The commuting diagrams he derived as part of his presentation gave an insightful foundation to underpin many of the issues that were covered during the week. His key claim was that map-generalisation must necessarily be informed by the ***purpose*** for which the generalised map will be used. This claim was supported/illustrated by the commuting diagrams he derived during his talk. Consequently he argued that "we give up the chimera of the general purpose map and concentrate research on mapping for particular decisions". This claim was warmly applauded by the "semanticists" in the audience, who confirmed the insight that semantics is often (perhaps always) informed by the ***purpose*** for which the semantics is to be used: meaning is always context-dependent.

Combining the previous two talks (a "context-dependent" position and the use of ontologies to model that context) was the starting-point for the presentation by **Sebastien Mustiere**. He argued that each geographic database reflects a certain point of view that can be be modeled as an ontology. He then proposed that such ontologies would be useful tools for analysing, selecting and comparing geo-databases.

Stephan Winter (all the way from Melbourne) focussed on the fact that spatial descriptions have widely varying granesize to which they apply, and reported on investigations to use results from semiotics to select the appropriate spatial grainsize for a particular information exchange.

Anna Zhdanova (Vienna) gave a very useful overview from an applications perspective, in this case the use of geospatial information for mobile services. She showed through a number of prototypes that in practice, end-users increasingly demonstrate the capability to successfully generate and exploit ontology items such as classes, subclasses, properties, instances, ontology mappings and rules.

Thursday April 16

First morning session (chair Marc van Krefeld)

On Thursday the first session started with short plenary reports on the four parallel breakout sessions of Wednesday afternoon (for more details see the appendix):
W1 Partitioning Large Datasets (reporting by Patrick Revell)
W2 continuous/gradual generalization (reporting by Peter van Oosterom)
W3 Cartographic and semantic aspects in webservices, incl. Mobile (reporting by Lars Harrie)
W4 Ontology of generalization operators (reporting by William Mackaness)

Dirk Burghardt elaborated on the EuroSDR project which tried to specify the generalization problem via 'cartographic constraints' (or better stated the 'optimization goals'). Basically there are two competing goals: legibility (needs space for clear presentation of features at smaller scales and therefore introduces deformations) and preservation (minimize change compared to the largest scale/real world representation). A taxonomy of different types of 'constraints' is used in elaborating on these generalization goals was presented. In the discussion following the presentation is was asked weather it is possible (or should be) to make a difference between hard constraints (must be met) and soft constraints (optimization goal) or if this just a matter of weighting. In the current setting this was not yet the case, but it could be imagined that this in indeed an important difference.

Next, **Martijn Meijers** presented the problem that in a vario-scale data structure the different polylines can not be generalized independently (without introducing topology errors). In the presentation also two initial algorithms were given (not yet implemented). In the discussion following there was one suggestion to look at a computational geometry paper avoiding topology errors when generalization lines in a topologically structured planar partition (most likely this was for two fixed scales, and not a vario-scale setting). It was also remarked that the lines causing most of the problems are related to linear features (roads, waterways), which are represented as area's on large scale maps, but should collapse to a single line form medium/small scale maps. Martijn responded that this was indeed an intended future enhancement for the vario-scale structure.

In the last presentation for the break **Lawrence Stanislawski** presented the problem of how to evaluate the pruning (part of the generalization) of an hydrographic network. He presented quality indicators to evaluate the resulting representation. Besides the pruning process these indicators also made clear that there where differences in the various regions (maps sheets) of the input data. A question whether the indicators did also influence the further improvement of the pruning process (in case dome weaknesses would be discovered) was answered by stating that this had not yet happened (until today). It was further emphasized that these indicators do for an important tools in specifying the resulting product and quality.

Second morning session (chair Rob Weibel)

After the coffee break **Anna Zhdanova** explained that the web, mobile, physical and virtual environments converge in one shared communication sphere. In this increasingly growing and getting more heterogonous environment, she emphasized the importance of the semantic technologies especially for specifying users policies and preferences the for mobile platform (but also for other web-terminal). One important aspect was to empower the user with capabilities to specify his/her preferences for various services (both using services, but also participating in creating services; e.g. by providing location information back to the network). In such a context the users would become also produces: 'prosumers' (of microservices).

Marc van Kreveld explained reason to treat the generalization of trajectories (represented by series of connected x,y,t-point) different from 3D polyline simplification: enable more efficient processing and analysis afterwards; e.g. looking for (sub-)trajectory similarities (especially difficult if start times are not equal). His simplification method was based on efficiently finding short-cuts and then to derive a more coarse representation which still reflects well the location and speed of the moving object.

Monika Sester presented a new approach to detecting buildings in Lidar data and performing building generalization. The input data starts segmenting the points in 'on terrain' (low) and 'off terrain' points (high). The building outline/boundary are detected (between the on and off terrain points). The result is a (wiggly) boundary with (far too) many points which should be simplified. Based on a random sampling technique straight lines are selected (in this 'redundant' boundary description). These separate lines segments are then connected again (taking building characteristics into account) by formulation the connection of the line segments as a least squares problem. Via parameter settings the amount/size of the selected line segments can be specified (which is then translated to more or less generalization).

Afternoon Session (Alexander Wolff)

The free time after lunch was well used for all kinds of planned and unplanned discussions/ meetings, such as a EuroSDR generalization meeting and meeting on a possible EU FP7 project proposal related to generalization.

The presentation and discussion part of the program resumed with **Julien Gaffuri** who proposed a kind of benchmark or test environment (input data and goal specification) for generalization compatible to 'The Cornell box' in 3D computer graphics. Care should be taken that this would include the relevant types of objects (geometries and thematic classes), sufficient range of scales (LoD), various types of space (urban, rural, mountain, coastal, etc.), types of generalization (model and cartographic), different user needs/goals/tasks, different data set sizes, different technical environments (open source software, web-services, mash-up), 2D, 3D, 2D+time, etc. In one of the questions after the presentations is was remarked that there is significant overlap between this proposal and the EuroSDR project (and it should be discussed how to combine these initiatives).

The 'webgen' initiative is a few years old and **Nicolas Regnauld** did now try to formulate some explicit advantages of this platform (or stated otherwise 'use cases for generalisation web services'). Some of the advantages are generic (reusability, clear wrapping of functionality) and some advantages do exist mainly for different user groups: researcher (evaluate existing tools/methods, built on work of others, 'publication' platform, get feedback), NMAs (design a generalization production line/rapid prototyping using various generalization services, but also implementing the new production line), and industry (discover new generalization algorithms, compare to own tools, publish demo's, make own system extensible with functionality from others). The future will learn how urgent these need are and in which form the web

The last presentation of the day was by **Claus Brenner** on "Meaningless" Maps or how can raw data (such as Lidar data or images) be used form certain tasks, such as orientation. Positioning, etc. In essence, somehow within this raw data it is attempted to detect the implicit feature by using knowledge related to these feature. Claus made the observation that getting to abstract representation of certain feature (is also the essence of generalization: abstraction). He further stated that some of these abstractions are for direct machine consumption (and not intended for human in the sense of map reading; often the goal of generalization). Finally he did give the following lessons: in interpreting raw data and also in map generalization one is looking for the high dimensional descriptors, that contain a lot of information (the essence, what set one thing apart for the others).

After the plenary afternoon presentations again four parallel breakout sessions where organized on the following topics:
T1 Pattern recognition in map generalisation (semantic enrichment)
T2 Semantics and computational geometry
T3 Ontology (semantics) of concepts at different scales
T4 Generalization for the machine (efficiency/real time aspects)

Demo Session

After dinner the group resembled for various plenary demonstrations (about 10 minutes per presentation). After all plenary demo's, more in-depth demo's followed in smaller groups of interested participants.

In summary the Thursday again focused on generalization and did not contain a lot of contribution related to semantic technologies or a mix of generalization and semantic technologies (beside some of the topics in the breakout sessions and some demo's). In reflection this might have been mixed a more.

Friday April 17

First morning session

On Friday the first session started with short plenary reports on the four parallel breakout sessions of Thursday afternoon (for more details see the appendix):
T1 Pattern recognition / semantic enrichment in map generalisation (reporting by William Mackaness)
T2 Semantics and computational geometry (reporting by Alexander Wolff)
T3 Scale-dependent ontology of urban concepts (report by Patrick Luscher)
T4 Generalization for the machine (report by Peter Hojholt)

Omair Chaudhry presented his research with William Mackaness on “Database Enrichment: Automatic identification of Higher order objects in large scale database”. He explained how functional sites, like hospital or schools, can be detected from detailed topographic data in order to make a semantic enrichment of the data. In order to do that, criteria defining the functional site should be formalised, by means of concepts related to the notions of connectivity, composition, shape or size among others. The work has been exemplified with results obtained on hospitals from Ordnance Survey base data.

Directly linked with the previous presentation, **William Mackaness** elaborated more generally on “Partonomic modelling to support context and reasoning in map generalisation”. He argued that the generalisation process needs to be task oriented. With examples based on journey planning, he pointed on the importance of notions related to partonomy to enrich spatial databases: a map is more made of functional sites than geometries, in term of cognitive concepts to analyse it. He then illustrated how this approach can be implemented with reasoning based on first order logic.

Werner Kuhn presented “Generalization of Geographic Information as an Ontological Problem”. He that there exist many types of generalisation, and that map generalization is a conceptual problem, before being a graphical one. As a direct consequence, ontology is a necessary underpinning for generalization. He then briefly presented existing ontologies of geographic concept and how those could be used and shall be enriched. He argued that we know very little about how conceptualizations are generalized and that there is not much work so far on (multi-)granularity in ontologies, opening the way to research on this topic.

Second morning session

In his talk entitled “Explicit storage of DCM instances considered harmful”, **Peter Van Oosterom** recalled that the distinction between DLM and DCM may be one of the most important contributions of our field. However, despite the general acceptance of the DLM-DCM paradigm, the explicit storage of DLM and DCM instances (and perhaps links between them) may be harmful. Several options could then be envisaged: 1/ only store map objects, which may be inefficient for spatial analysis, 2/ store both objects of DLM and DCM and links to propagate updates, 3/ only store objects from DLM while slightly adapting them for a default efficient visualisation. He argued that the last option that may theoretically hurt at first sight may be practically clever as well as theoretically justified. He also wondered what happens with the DLM/DCM vision with the ever stronger the ‘per-theme’ trend in GI domain (e.g. INSPIRE).

Leila de Floriani gave a talk on the ”generalization of Morse complexes in arbitrary dimensions”. She described the Morse Theory, a mathematical theory for modelling morphology of scalar fields, based on the notions de minima, maxima and saddles. Such morphologic models describe the field in a more effective and compact view than classical geometric models, which is important for analysis, exploration and analysis of fields. She then presented generalisation operators in this model, which have among others the interesting particularity of being reversible.

The last talk of the seminar was given by **Esteban Zimanyi** on multicriteria decision analysis (MDCA), which can be seen has a set of functions to optimise, while no global optimum exist and a compromise shall be found to determine the best solution to a problem. He presented how GIS and MDCA tools can be integrated. He illustrated the approach followed by the GAIA tool, based on the notion of flow as the intensity with which an alternative is prefered / not prefered to the others.

Monika Sester concluded the seminar with the main insights she learnt from it, and asking participants their feedback.

As a kind of conclusion…

The idea of bringing together researches from various community, from semantic web specialists to geometry specialists has been widely thought of as fruitful. This led to many discussions on links between those fields of research. Some particular issues to be tackled appeared in many questions and discussions, and may lead to further research, like:

- How cognitive and semantic aspects could be integrated in the generalisation process that is (may be too much) geometry-oriented?
- How to handle the notions of fuzziness or uncertainty in semantic web techniques?
- Shall generalsiation bemore task-oriented? How to do that?
- How to handle the notions of time and updates in both domains?
- ....

APPENDIX: REPORTS OF BREAK OUT SESSIONS

W1: Partitioning Large Datasets

Report by **Patrick Revell**

Why do we want to partition?

- **Computational reasons**. It is not possible to process entire dataset at once so need to break it down into manageable units – "divide and conquer" approach.
- **Maintain characteristics**. Different regions of the dataset may have different characteristics that require different generalisation approaches. Identifying these regions allows suitable processes to be applied depending on the region.

The best method of partitioning varies from theme to theme (and may even vary from operation to operation within that theme).

There are two types of partition:

- **Regular**. A regular grid is applied to the dataset. Each grid square is processed with a small overlap with the surrounding grid squares. The features around the edge are discarded since they have been generalised out of context (so some features are processed more than once). A similar approach is to use an arbitrary tessellation of space, such as county or state boundaries.
- **Irregular**. These are derived from geographic information that is explicit or implicit in the source data. It is important to distinguish between features that can change and features that provide context (i.e. the partition boundaries).

The discussion concentrated mainly on irregular partitions and the techniques used for creating them:

Trans-Hydro Graph. This is created by taking road, track, railway, river and canal networks, intersecting them, and forming partitions from the closed loops. Some agencies also include the boundaries of large forests when forming such partitions.

Urban-Rural Areas.

- Urban areas can be identified by buffering buildings and merging the buffers together (Boffet, IGN France).
- The problem with this approach is that ribbon developments around the edges of cities get included in the urban areas. A better solution takes density into account. To do this, for each building find the n nearest buildings (eg. 20), sum the distances, make the buffer size inversely proportional to the sum of the distances, then merge all the buffers (Chaudhry, University of Edinburgh).
- It is computationally intensive to merge together thousands of buffers, so an alternative approach is to buffer road junctions instead of buildings. The initial boundaries can be refined using building density buffering and finally by adapting the boundaries to follow topographic features (Chaudhry, University of Edinburgh).

The AGENT project demonstrated that the trans-hydro graph is good for generalising buildings in urban areas. For rural areas the phenomena are more dispersed, so for these it is better to cluster together groups of features that are in conflict (Duchene, PhD, IGN France).

Mountainous Areas/Plains. These can be computed from the DTM. It is possible to use a coarse DTM to obtain "rough and ready" partitions, that can then be refined using a more detailed DTM (Chaudhry, University of Edinburgh).

If networks are being used for partitioning, how do we generalise the networks themselves?

Hydrography. USGS is working on generalising the National Hydrography Dataset for the USA, which is gigabytes in size (Stanislawski). This is broken down into regional watersheds, sub-regions and sub-basins for processing. The issue with this dataset is that it has been collected with a wide range of different capture standards, so density is not consistent across the dataset. A density buffering approach could be used to identify regions of different densities, but the “expected density” would still need to be determined, perhaps by detecting channels in a DTM.

Roads.

- A simple approach is to use the main road network (Motorways and A Roads) to form partitions, then generalise the roads inside these partitions. However there are issues with maintaining consistency across partition boundaries and you still need to find a way of generalising the Motorways and A Roads.
- Another idea is to generalise roads in rural areas first, identify the entry points into urban areas, then generalise the roads in urban areas taking these entry points into account (Touya, IGN France).
- For road displacement a good solution is to identify groups of conflicts that can be solved together. This has been implemented using flexibility graphs at IGN France and a similar approach has been developed at Ordnance Survey (Thom).

The remaining part of the discussion covered any other themes that may require partitioning:

Landcover. The trans-hydro graph can be used for generalising landcover, but it is important to check the results are homogenous across partition boundaries. For example large forests may be divided into many small pieces by the trans-hydro graph. One approach to maintaining consistency is to use Macro agents to monitor generalisation of the dataset at a global level. Work on identifying forest boundaries has also be carried out by Chaudhry (University of Edinburgh).

Text. A regular grid can be used for placing text. This approach has been implemented in the Maplex text placement software from ESRI.

Coastline. The coastline is not normally large in data volume, but it can be formed of sections with different characteristics. Perhaps coalescence can be detected, then parts with similar characteristics can be clustered together.

Some further discussion points not covered in the break out session:

- **Handling updates**. Partitions can be used for managing updates – i.e you just re-generalise inside the partitions that have changed. But what happens if the features forming the partition boundaries themselves have changed?
- **Partitioning for data matching**. How to choose partitions when matching between two inconsistent datasets, eg. buildings eg. road networks.

W2: Continuous/gradual generalization

Report by **Peter van Oosterom**

Continuous or gradual?

Continuous generalization is a term that implies smooth transformations when changing between scales. However, in reality it are often small steps (much smaller than the fixed number of scales now typically produced my NMAs). Continuous (or smooth) zooming may the use these smallest steps in combination with 'display' techniques such as morphing, shrinking, fading, etc.

Purpose of gradual generalization

1. user interface aspect: smooth zooming (continuous)
2. data management aspect: non redundant data storage (compared to multiple-representations)
3. data communication: progressive transfer of refinements

State of the art

The current products, state of the art (e.g. Google Maps) perform quite well and do give the users already a bit of the continuous generalization feeling. Is more research needed to further improve the situation? A discussion followed and it was confirmed that there is still (sufficient) room fro improvement compared to the 'brute force' approaches (e.g. having 15 explicit different scale representations). A big challenge will for combining different themes (or user generated/added content). How can the result still be a consistent representation at all required scales (LoD's)? Also different users might have different preferences for the different themes (this could then be specified via theme/object class sliders indicating the relevant of the different themes). The resulting continuous scale representation should then be created based on user preferences (and for all scale levels representation should be consistent).

Terrain modelling

An example were two different models should not result in inconsistent representations was given in the context of terrain/elevation modelling. One approach is to have a raw point/triangle based representation which can be queried at different scales (and the appropriate number of points/triangles is selected). Another approach is modelling the critical points: peaks, pits, saddle points and base the generalization structure on the merging of the neighbour critical points (e.g. two peaks and a saddle points are merged to one peak) and storing the resulting structure in a graph. Even if this might be considered the same theme (elevation), keeping the two multi/gradual scale representations consistent is an issue.

Procedural approach

An approach that was suggested in order to enable non-redundant representations is to apply more procedural techniques. This should in principle work in moth direction: from detailed a procedure to coarse (e.g. a procedure to collapse two neighbour buildings) and from course to detailed (in this case the procedure does need some 'delta' information as input parameters). Especially in case of reoccurring patterns; a procedural approach may be efficient (as in the PostScript printer language).

Type of data to be generalized

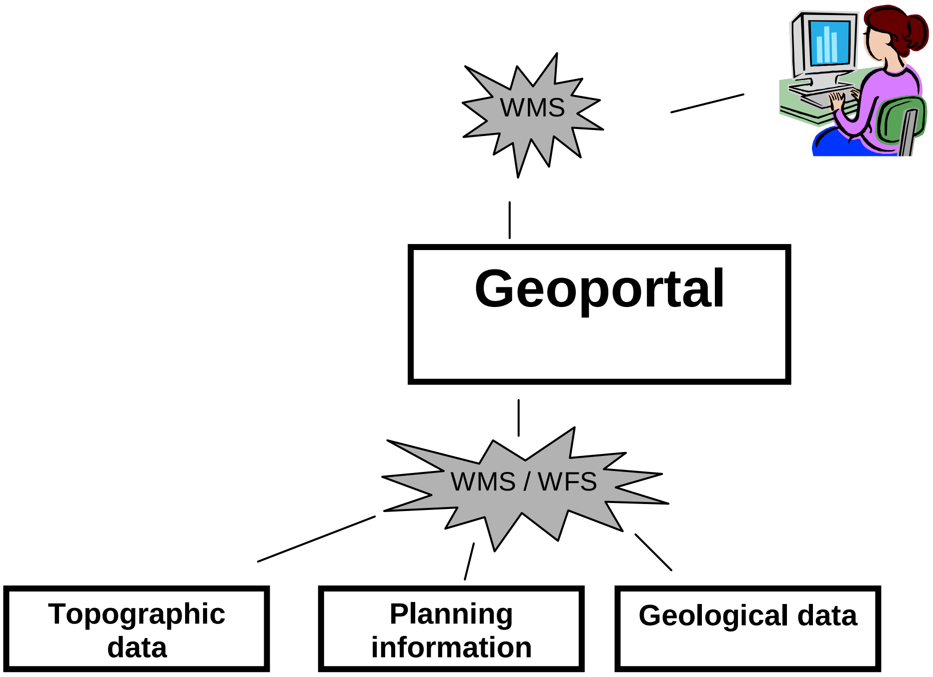
Does it matter for gradual generalization if the theme is related to discrete (e.g. man made objects; roads, buildings) or more continuous themes (often natural; e.g. elevation, vegetation, etc.)? The answer to this question was not that clear. However, it is obvious that there are differences; e.g. currently a set of 10 building polygons may be generalized to 3 building polygons (on the next smaller scale), then to one built up-area polygon (next smaller scale) and finally represented by a point (on the smallest scale). How to deal with this in context of continuous generalization?

W3: Cartographic and semantic aspects on web services

Report by **Lars Harrie** and **Heiner Stuckenschmidt**

Background

Several countries are currently working on setting up geoportals as part of their national spatial data infrastructure (SDI) (and this is also a requirement of the Inspire initiative). A key ability of these geoportals is that the user should be able to view (and download) data from several sources from one access point. This will certainly make the access to geospatial data easier. However, there are also cartographic and semantic challenges that have to be solved. In this discussion group we discussed some topics concerning both download services and view services (as in the figure below) and some possible solutions.



Download services

Problems can arise when the user uses different download services to combine data from different, possibly heterogeneous sources and wants to combine them into a single view in the geo-portal. In principle, the problems that can occur in this situation are the typical problems of data integration that can be found in any domain, i.e. inconsistency, redundancy and differences in granularity and conceptualization. In the context of geo-information, these general problems manifest themselves as follows:

1. Inconsistency in data + redundancy in data: In some cases several organizations have same object types (e.g road data is stored both by NMA and road administrations). Then inconsistency / redundancy of data could be a problem due to e.g. different update cycles.

2. Different Levels of Detail: Organisations are using different level of details in their data, which will cause problems when the user is merging data from different sources.

3. Different classifications: Organisations are using different classification schemas which of course are problematic when merging data.

View services

View services have the same problem as download services, but come with some additional problems. In particular, in the case of download services, the way, information is represented in the portal was defined by the user accessing the different sources who could ensure a uniform representation. In the case of view services, not only the data, but also the way it is represented is defined by the local sources leading to the following additional problems:

4. Symbology: The base data (e.g. topographic data) is often rendered to optimize its own visualization. This is problematic when other data (additional data) is put on top of the base data. E.g. there could not be any suitable color left. Another problem is that different servers might use the same symbols for different object types.

Possible solutions: The user must be given the possibility to change the symbology in the view services e.g. with SLD (styled layer descriptor). Another possibility is that the user can choose among a set of symologies (stored e.g. on the server or geoportal level).

5. Overlapping problem: Data can overlap or be too congested for cartographic visualization.

Potential Role of Semantic Technologies

It has been widely acknowledged, that semantic technologies can play a major role in overcoming data integration problems in different domain. Research in the database as well as the semantic web area have developed technologies for describing the intended meaning of data in different sources and using these definitions for defining semantic relations between different sources that can be used to integrate information in a meaningful way. Thus a natural question is whether semantic web technologies can also help to overcome the problems we identified above. In the following, we discuss the different problems and identify the potential contribution of semantic technologies to solving the problem. As we will see, some of these problems can naturally be addressed using semantic web technologies while others elude a solution involving semantic technologies.

1. Inconsistency and redundancy in data: Checking consistency of definitions is a basic functionality of semantic web technologies and can be implemented using the web ontology language OWL [Horrocks et al 2003]. This, however, requires the data from the different sources to be described as instances of a common ontology. This ontology has to specify explicit consistency constraints for the data and is limited to certain types of inconsistency. In particular, semantic technologies can only be used to identify conceptual inconsistency such as legal combinations of types the same object can have or legal types of objects it can be in a certain relation with. Other types of inconsistency resulting, for example from outdated data cannot always be found as semantic technologies cannot check data against the real state of the world, but only against data from another source. Checking redundancy is not directly supported in OWL.

2. Different Levels of Detail: Levels of details are a problem that is very characteristic for geo-data. While the granularity of data is also an issue in other domains, the issue of granularity has some very specific properties in the domain of geo-services. Here the issue of granularity is a geometric rather than a semantic problem. If the data from two sources do not have the same level of detail, the more detailed map has to be abstracted. This can be done using existing abstraction methods that are mostly geometric by nature. It is not entirely how semantic technologies can help in this case.

3. Different classifications: Using semantic technologies to integrate heterogeneous classifications used by different data sources is one of the more promising applications of semantic technologies in this context. In fact the use of semantic technologies for integrating heterogeneous object catalogues has already been described in the literature [Stuckenschmidt and van Harmelen 2004] and there is a rich literature on matching ontologies [Euzenat and Shvaiko 2007] that can also be applied to object catalogues. A problem not adequately addressed by current semantic web technologies is vagueness of concepts in the geographic domain (what is a mountain as opposed to a hill?). Recent work combining OWL and fuzzy reasoning [Straccia 2005] addresses this problem to some extend but so far, this extension is not an official language and there are no experiences with using Fuzzy OWL for modeling geographic concepts.

4. Symbology: the problem of in compatible symbology that arises from the use of different view services cannot be resolved using semantic technologies; however, it could be possible to use semantic technologies for detecting problems with symbology that might not be spotted by the user at first sight. In particular, it is possible to build a semantic model of types of objects shown in the geo-portal. The description of such concepts could contain a relation that links the type with the symbol used to represent objects belonging to that type. In order to make sure that no symbol is used for different concepts, this relation could be specified as being one-to-one. When a new data source

is included in the portal, it needs be represented as an instance of that model. If a symbol is now used for different object types or the same object type is represented by different symbols, this causes and inconsistency in the semantic model that can be reported to the user.

5. Overlapping problem: This problem again is a typical example of a problem with no obvious use of semantic technologies. It might be possible to also build a semantic model of the data that constrains the configuration of polygons on the screen in such a way that potential problems result in inconsistencies in the semantic model. Such a model, however, will be mainly concerned with spatial constraints that cannot easily be encoded in semantic web languages. This makes it unlikely that semantic technologies are a good choice for solving this particular problem.

Conclusions

The integration of multiple spatial data sources into geo-portals comes with a number of potential problems related to mismatches in the data to be combined in the portal. We distinguished between problems related to download services and additional problems arising from the use of view services where conflicts can also include the representation of spatial objects. As briefly explained, some of these problems could be addressed using semantic technologies. In particular, semantic technologies can be used to detect inconsistencies across data sources provided that the data from different sources is described using a common semantic model. Further, semantic technologies provide support for the integration of heterogeneous classifications of objects which is a fundamental requirement for having a meaningful integration of different datasets. Beyond these possible applications there are also a number of problems like different levels of detail and overlapping of polygons that should not be addressed using semantic technologies but should rather be addressed using computational geometry and related methods.

References

Ian Horrocks, Peter F. Patel-Schneider, Frank van Harmelen: From SHIQ and RDF to OWL: the making of a Web Ontology Language. J. Web Sem. 1(1): 7-26 (2003)

H. Stuckenschmidt and F. van Harmelen. Information Sharing on the Semantic Web. Advanced Information Processing, Springer Verlag, Berlin, Heidelberg, 2004

Jérôme Euzenat, Pavel Shvaiko. Ontology Matching. Springer-Verlag, Berlin Heidelberg (DE), 2007.

Umberto Straccia: Towards a Fuzzy Description Logic for the Semantic Web (Preliminary Report). ESWC 2005:167-181

T1: Pattern Analysis in Map Generalisation

Report by **William Mackaness**

The group discussed the need for pattern analysis in map generalisation and explored different approaches to solving the pattern analysis process.

Why do we need pattern analysis

We need pattern analysis tools in order to:

1. trigger a set of generalisation operators (fitting a solution to a specific constellation of features)
2. to make explicit the semantics implicit in pattern (for example that linear, regular, or random patterns tells us about the feature we are looking at).
3. to guide the evaluation process where there is a specific ‘arrangement quality’ that we wish to keep after generalisation (such as manhattan space quality).

What do we mean by pattern analysis

The pattern inherent among an arrangement of features (gestaltic properties).

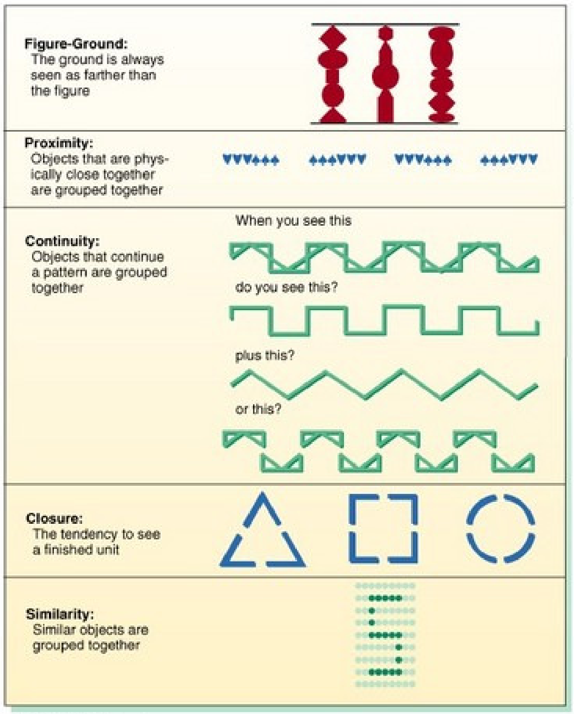


Figure-Ground:
The ground is always seen as farther than the figure
Proximity:
Objects that are physically close together are grouped together
Continuity:
Objects that continue a pattern are grouped together
When you see this
do you see this?
plus this?
or this?
Closure:
The tendency to see a finished unit
Similarity:
Similar objects are grouped together

© 2000 John Wiley & Sons, Inc.

Figure 1: gestalt as the building blocks of patterns
http://img.photobucket.com/albums/v218/terrymockler/Gestalt.jpg

Hierarchy of patterns

Discussion highlighted the fact that we see hierarchies of patterns. At one level we see a pattern among a set of regularly spaced buildings. At another scale we might see repetition of that regularly spaced pattern – say across road partitions. (akin a little to fractal space). So any pattern recognition technique needs to take this into account.

Approaches

Can we exhaustively list all possible patterns – and create a library of patterns (anthropogenic and natural), and use the library to search the data for specific instances of those patterns. – ie a classification of the map in terms of its patterns?
OR
Can we use unsupervised learning techniques to identify patterns (and thus not worry about any library of predefined patterns).

What sorts of ideas might help us with this problem?

Computational Geometry: work already done using MST.
Computer vision: Scale-invariant feature transform (or ***SIFT***) (an algorithm in computer vision to detect and describe local features in images) could be applied in this context.
Computer vision: face recognition technology is standard in cameras. Can this be adapted for recognising specific patterns?
Machine learning techniques: given that we have lots of detailed maps and their generalised equivalent, surely there is an opportunity for a machine learning approach?
Remote sensing: Can we use RS techniques such as fourier transform?
Handwriting technologies: is there a parallel between pattern recognition in hand writing and patterns in maps?

What do you do once you have the pattern?

Patterns afford a route to semantic enrichment. Potentially the feature now knows how it contributes to a particular set of patterns – in other words it knows something of its context. Pattern analysis is but one form of enrichment, that could help you describe or characterise a place (Figure 2).

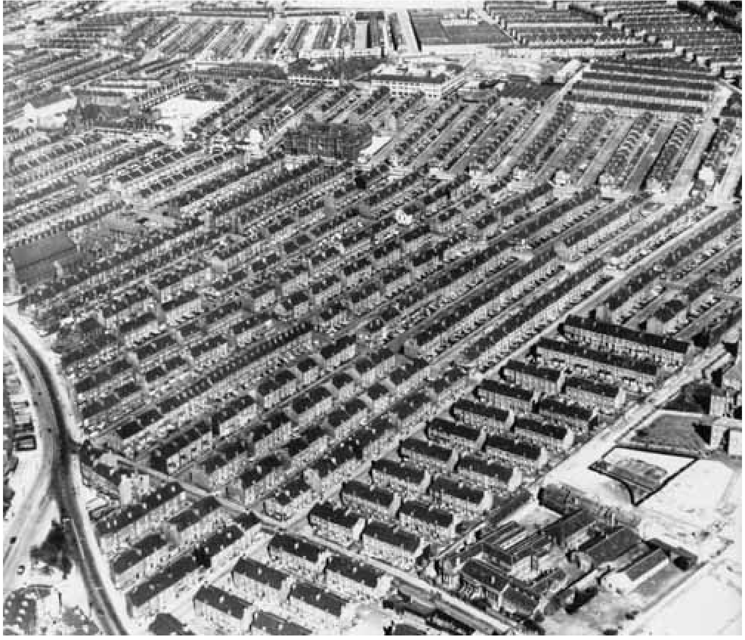


Figure 2: Characterising spaces: http://www.leodis.org/imagesLeodis/screen/46/2007731_164346.jpg

With such information we might be able to answer questions such as ‘is this a nice place to live?’. Being able to characterise a region is key to its immediate recognition (Figure 3).

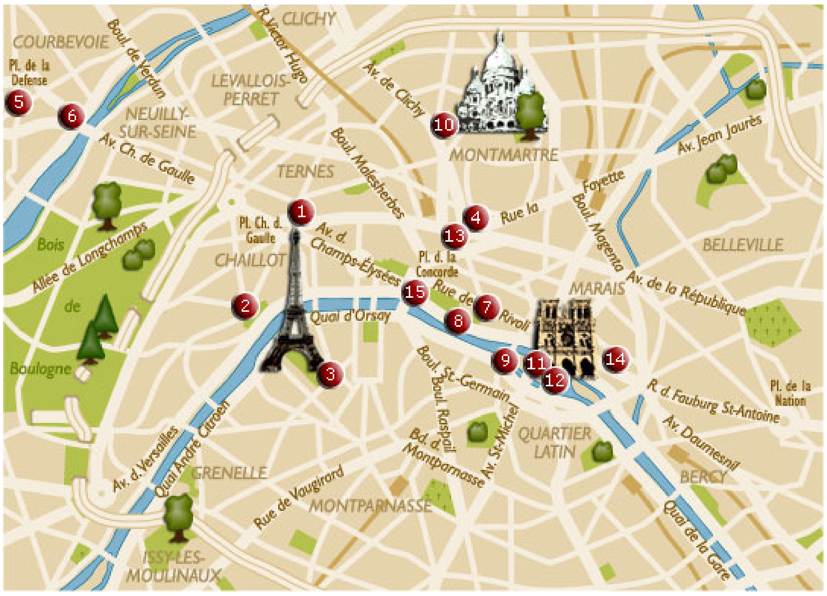


Figure 3: Where are we now? http://www.stadtplandienst.de/objects/euro360/paris.html

Conclusion

Pattern analysis is a very important technique and can support a range of activities related to multi scale view of geographic information.

T2: Semantics and computational geometry

Report by (**Peter van Oosterom and) Alexander Wolff**

(Miss) Match between semantics and computational geometry

Somehow there is the strong feeling among the participants of this break-out session that both formal semantics (theories and tools) and computational geometry (theories and tools) are needed for solving the map generalization challenge. However, the two theories and tools are currently quite far apart: computational geometry needs well defined (geometric) problems as input and a small rephrasing of the problem statement may lead to a complete different problem and solution. At the opposite side, formal semantics tries to function in an environment which is less well defined (and perhaps sometimes even containing some non-compatible facts).

Added value of the combination

It was argued that with the help of formal semantics we may try to develop more flexible solutions, that is, not hard code everything (which computational geometry solution to apply to which objects/situations), but use some generic intelligence to characterize this and apply the right computational geometry solution. This may be related to characterizing the problem area (rural or urban environment) and some (hierarchal) classification of object types (top level: natural – man-made and then further refinement), which may be beneficial in the decision process based on formalized knowledge, which solution to apply. Also the users wishes/requirements might be formalized using semantic web technologies (and therefore be applied in a more flexible solution). Once these characteristics have been defined, we might use computational geometry tools in two situations: 1. to analyse a specify data set (collection of instances) and attach to this the proper characterisation (including more complicated patterns) and 2. after having characterized the situation and knowing the user requirements, the reasoning process (based on formal semantics) could figure out which computational geometry tool to apply in which situation (and in which order).

How to make this happen?

The two worlds are currently quite far apart, and despite the above described complementary value of the two approached, it was not yet crystal clear to the participants of the break-out sessions, how the benefits could be obtained in practice (by realizing some kind of hybrid system). This system would then contain the collected generalization knowledge and could be considered as an ‘automated designer’ for the development of a generalization process in a specific situation (given data sets and user preferences). Or perhaps more modestly stated: the orchestration of the workflow within the generalization process. The group could not get beyond the advise: ‘more research is needed here’.

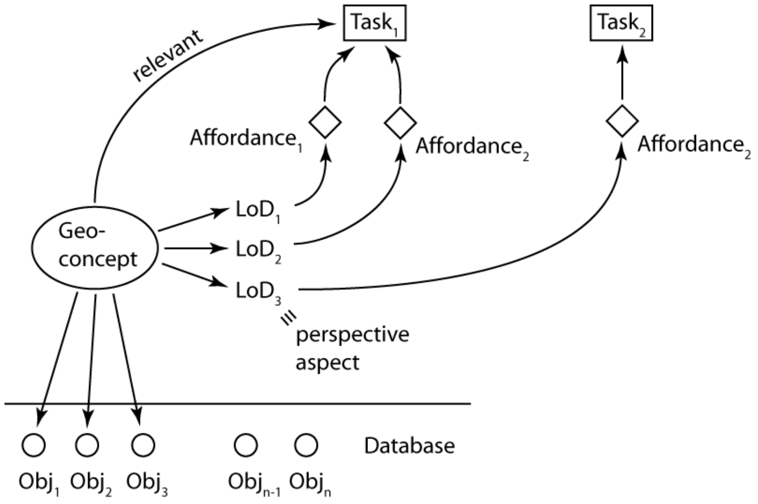
T3: Scale-dependent ontology of urban concepts

Report by **Patrick Luscher**

Task-oriented portrayal of information

Adaption of maps to user-specific tasks requires understanding of the concepts involved when humans solve specific tasks. Concepts differ from patterns in cartographic databases in the sense that concepts are represented by a specific setting of objects, but the same setting can denote different concepts depending on the task. An example is as follows: A railway serves as a way of connecting places for someone travelling by train, but as an obstructing barrier for a pedestrian.

Hence, concepts are a means to connect user-tasks with objects in cartographic databases.



Scale-dependency
Can we assign a certain scale to concepts where they are important? One example which was discussed is hierarchical navigation, where at the local part a gate at the front of the house is important, whereas, when travelling within a city, types of districts are important, and, when travelling between cities, a location representing the whole city might be sufficient.
Hence, the task is a way to describe the level of detail. Indeed, many map-users identify what they need to see in a map with map scale.

Affordance
Concepts are relevant to a specific task if they afford the task, such as "this kind of feature would afford hiking if it were included in the map." It was also discussed that a concept is relevant for a task (and not its representation), i.e., there exists a direct relation between concept and task, and then a representation affording the task has to be found.

T4: Generalization for the machine

Report by **Peter Højholt**

Reason to generalization for the machine:
Maps can be too large or not suitable organized for a machine to handle them efficiently

What the end machine does:
From the generalized map the machine makes an analysis and provides an action/answer that is either not a map or is another map meant to be used by a machine. The sole purpose of the machines analysis is ***not*** to make a display map.

The discussion of the Partitioning of a data set is an example of a problem that is only interesting when the data at hand is not suitable for the machine. The purpose of the partioning is to generalize data in such a way that it becomes more useful to the machine.

Machine capabilities:
von Neumann machine:
Fast but not infinitely fast
Storage is large, but not infinite
Broad bank (or no?) internet access

What we want:

Data that are organized such that map-problem solving occurs in lon(n) or at most n*log(n) time
How do we serve up our data for the machine to work most efficiently?

Compared to generalizations made for humans:

Things and concepts that we don’t need:

Scale
Displacement (harmful) (not gene)
Enlargement (harmful) (not gene)

Not a problem:

Local density of data

Suggestions of granularity of relevant problem sizes to discuss:

The whole world	(weather forecast, environmental studies)
Continental/Sub-Continental	(Emergencies, Flooding, Pollution)
Local (country) size	(Transport)
Very local size	(Noise)

Suggestions of granularity of relevant machine sizes to discuss:

Multiple servers (grid)
Server
Workstation
Hand held