Grant application for RGI-232

Project title (NL): GeoInfoNed, een multimedia geo-database infrastructuur

Project title (UK): GeoInfoNed, a multimedia geo-database infrastructure

Management summary
Database management systems have become the nucleus of many spatial information systems. They provide the infrastructure for both the storage of spatial data and associated administrative data to aid the user in his quest for valuable information.

The landscape for Geo Information is however changing drastically: Next to simple text and numeric attributes, it becomes mandatory to seamlessly integrate more complex types of data. This can range from multimedia data (video streams from surveillance cameras), via aggregated sensor data to highly structured XML data. To overcome information overload a user needs a filter and analysis step that aggregates the staggering amount of data into meaningful features.

Currently no generic Database Management Systems are available that provide the performance and scalability needed for such a system.

The GeoInfoNed project brings together two internationally leading research teams to develop a reference platform for multi-media enriched Spatial Database Management System using GML as an interface. A spatially enhanced XQuery implementation provides a stepping stone to link into the emerging visualisation tools, e.g. Google Earth.

Leading party

Organization: TU Delft – Onderzoeksinstituut OTB Based in: Delft (Netherlands)

Name, title and initials of officer with authority to sign: prof. dr. P.J. Boelhouwer, managing director OTB research institute

Legal form: University

Chamber of Commerce no.: registered in:

Leading party is *(enter x as applicable)* X registered for VAT 0 exempt from VAT

Bank account number: 54.4385.543

Bank: ABN-AMRO

Contact person

Name: Wilko Quak

Complete only if the contact person is not yet on the RGI distribution list:

Position: Assistant Researcher
Telephone: +31-(0)15-2783756
Fax: +31-(0)15-2784422
E-mail address:C.W.Quak@tudelft.nl

Project consortium

1. OTB/Tudelft/Sectie GIS Technologie — implementer
2. CWI Amsterdam/Cluster Information Systems — implementer
3. Rijkswaterstaat/AGI (RWS) — user
4. CycloMedia — user

Social relevance *(max. 1000 words)*

- *State which obstacles and/or challenges in social and/or economic issues are being addressed.*
- *Provide details of the social and/or economic benefits of executing the project.*
- *Provide arguments for why the project cannot be carried out without a Bsik grant.*
- *State how the research strategy takes account of the requirements imposed by the application area on the envisaged knowledge and technology.*
- *Demonstrate that the project objectives are supported by those with a demand for the knowledge, or end-users.*
- *Demonstrate how the results of this project can be generalized for use in other situations*

Database management systems have become the nucleus of many modern Spatial Information systems. They provide the infrastructure for both the storage of spatial data and associated administrative data to aid the user in his quest for valuable information.

The landscape for Geo Information is however changing drastically: Next to simple text and numeric attributes, it becomes mandatory to seamlessly integrate more complex types of data. This ranges from multimedia data (video streams from surveillance cameras), via aggregated sensor data to highly structured XML data. Since the emergence of GML/XML as an exchange format for geographic information more and more information will be available for analysis and integration.

In order not to drown in the sea of information, a user needs a filter and analysis step that reduces the staggering amount of data into meaningful features.

Many geo-information analysis tools are built for a specific environment and cannot easily be integrated with other analysis tools. In the GeoInfoNed project we plan to develop a platform for the integration of data and tools. Although a single technological solution is beyond the research horizon, we expect an ever growing collection of geographic and multimedia datasets and their analysis programs to be integrated in a generic way.

GeoInfoNed is organized around a growing number of participants, who bring in real-world cases and challenges. It results in an experimental driven research methodology, which is expected to produces tangible results quickly. By including project members from industry with different backgrounds and size, structure and interests, we maximize the functionality of the reference implementation and provide the widest view on potential commercial interesting applications.
One of the prime results of the project is the first GML/XML enhanced open source spatial DBMS. This result can be used in other situations in two different ways:

- Businesses, developers, and researchers gain ease of access to a state-of-the-art and free software platform.
- The software comes with no strings attached; it can be refined and commercialized quickly.

Partners are in the front seat of this development, because their data and their challenges are the focal points of research and development. Although the program is technically high risk, a partial success and the dissemination through the open-source license model guarantees a quick take-up.

At the end of the GeoInfoNed project we plan to have a system that can handle questions illustrated below. Please note that these questions are: 1.very hard to answer with current DBMS technology. 2. are of serious interest for society and 3. all data needed for the answer can be provided by the project members:

- **Datamining.** Analysis and prediction of road accidents. By linking existing accident registrations to other datasets (weather information, socio-economic information, etc.) causes for accidents can be detected and handled.
- **Monitoring.** Detect the pavement type and quality of all roads by performing image processing on the CycloMedia images. This information can be used in behaviour analysis of cyclists.
- **Planning.** Detect the amount of space for parking cars: Parking lots are well mapped in the Netherlands but nobody is registering the non-official parking places.
- **Monitoring.** Find out whether the colour of buildings near highways is significantly polluted because of black smoke emissions from cars.
- **Tracking.** To track behaviour of vehicles over a road network

Scientific quality *(max. 1000 words)*

- *State which scientific issues the project addresses*
- *Summarize the state of the art in the discipline underlying the project.*
- *Provide arguments for the innovative content of the envisaged technology and/or scientific research.*
- *Provide arguments for selecting the given R&D approach rather than alternatives.*
- *State the extent to which the selected approach will lead to the development of the envisaged knowledge and/or technology and/or skills.*

- *State the contribution made by the project to RGI's research ambitions.*
- *Indicate whether and, if so, how the project is compatible with European research programmes, such as the EU's sixth framework programme.*
- *Indicate whether and, if so, how you cooperate with international partners or use international databases.*

The real-word cases provide a basis for scientific progress in 1. **Heterogeneous databases,** Integration of very diverse data sources in one DBMS. 2. **Spatial querying.** Execution of complex analysis tasks inside a DBMS.

Integration of very diverse data sources in one DBMS: The project plans to bring the storage and retrieval of spatial data mixed with any other type of data at a higher level. The GIS community has recently moved from file-based systems to centralized DBMS systems. This has moved the geo-information more to the core organizations and has increased the awareness of the valuability of geo-information.

Spatial support in the open-source database community is still meagre. Currently, PostgreSQL has a decent support for spatial data (with PostGIS). It is, however, tuned towards support of the relational model and SQL. Previous research has convincingly shown that the impedance mismatch between spatial data and the relational model is still large. Commercial product seem to compete in stability and not in performance.

Execution of complex analysis tasks inside the DBMS: Many complex operations on spatial data are currently performed in the following way. 1. The data is copied from the database to a specialized program. 2.The data is processed. 3. The results are copied back to the database. This elaborate route is chosen because the operations provided by the DBMS are not sufficient for the efficient execution of the complex operations. GeoInfoNed plans to solve this problem by making an extensible DBMS platform with efficient operations that allow the analysis to run inside the DBMS.

The emerging XML-based data model with its query language XQuery provides a much more natural descriptions for the spatial objects. However, it requires a DBMS that understands XML storage and can handle XQuery. Both components are already available in the MonetDB, however, the spatial XML reasoning component is missing. Early research and comparison of Monet and Postgres against geo-information systems has already proven the performance superiority of the MonetDB:

- PostgreSQL has a traditional DBMS architecture that is not very suited for handling various types of data (XML, video streams…), the architecture of MonetDB has proven to be very efficient in handling these types of data.

Another consequence of the fact that many spatial analyses are done outside the DBMS is that this makes it very hard to integrate the different analysis tools or migrate an analysis tool written for a specific problem domain to another domain. By providing more complex operations in the server this problem will be lessened.

The GeoInfoNed project brings together two research groups working at the heart of the matter. The group in Delft is well informed and active in the area of modelling geo information as it pertains to end-user applications. Together with the end-users they can select, encapsulate and research different approaches to generate meaningful GML/XML documents and prepare them for database management.
The group at CWI is a leader in database technology. Its XML database engine is one of the fastest systems known and also scales to very large XML documents. Exposure to the

domain of GML/XML stresses the research envelop in areas such as efficient updating of XML stores and also calls for the spatial query primitives in Xquery.

Plan of action *(max. 2000 words)*

- *Objective: Describe the project objectives and state the scope of the project*
- *Activities and division of tasks: summarize the activities to be performed and show how the activities are divided among the members of the project team.*
- *SMART result (SMART: Specific, Measurable, Acceptable, Realistic, Time-framed). Summarize all intermediate and final results and tangible deliverables to be produced during the life of the project. It is important that the emphasis is on results (output, nouns) and not on activities (input, verbs).*
- *Quality assurance: describe the measures to be taken to guarantee the quality of the process, the results, the deliverables and the dissemination of knowledge.*
- *Planning and phasing: present a clear plan stating at least the following: commencement date, most important milestones, go/no go decision points, the critical path, delivery dates of intermediate and final results and total elapsed time. The minimum planning detail is quarters per calendar year.*

Objective:

The GeoInfoNed project plans to build an open source reference implementation for a Spatial DBMS in which a big number of sources of spatial and multimedia/XML data can be integrated. Querying this source of diverse information shall be done with extension of the query language XQuery that will be linked to dynamic web portals via modern web-technology.

As a starting point for the development of the DBMS we will use the Dutch open source DBMS MonetDB developed at the CWI. Its legacy spatial module proofed to be the fastest spatial DBMS at its time (see key publications). Other modules available in MonetDB include support for multimedia types and support form native XML storage.

As a starting point for the use of data we will integrate the datasets of RWS and Cyclomedia in this DBMS supporting queries that span across the datasets of both parties.

The bulk of the work will be done by two post-doc researchers during the two year that the project runs. One will be posted at the CWI and the other at the section GIS-technology at the TUDelft. On both sites they will work closely together with experienced research staff, this ensures an easy access to specialized knowledge that is available at the CWI and TUDelft.

Scope:

The project will extend the existing DBMS platform MonetDB with GML/XML spatial functionality. Multimedia and XML functionality is already available in the experimentation platform. It provides the stepping stone to realize the envisioned prototype within the timeframe foreseen.

Consortium composition

The partners of the consortium where carefully choosen.

- In research a two-way transfer of knowledge is most beneficial. Each participant benefits from a strong interaction with others. Both scientifically and to prepare for new, economic viable services.
- CWI has a strong reputation with respect to high performance database architectures and wish to proof their technology in as many fields as possible.
- OTB/TUDelft has a vast knowledge on current spatial database technology and wishes to leverage that knowledge in a novel DBMS architecture with a diverse collection of spatial data.
- CycloMedia is a commercial company with a huge dataset. Their data becomes extra useful when linked to other datasets. They are keen to investigate the scope and opportunities of GML/XML as a basis for this integration.
- RWS/AGI is migrating to an organization where its data is stored in centralized servers. For this process they need detailed knowledge of what is possible and what is not possible. Collaborating in a research project is an excellent way to build up this expertise.

Project Planning

The project will last for two years (Q1-Q8) and is split into 8 work packages (WP1-WP6).

WP0: Project Management and coordination (Q1-Q8) (TUDelft)

- Coordination of all project activities, set up website; Organization of meetings, incl. reports and minutes (internal and external); Progress control; Tuning between packages; Providing the central administration; Contact point (status reports and financial overview) between the project and the program office of RGI

CWI:

- **WP1**: During the first half year, CWI will upgrade its old GIS module developed for MonetDB Version2 and make it compliant with state-of-the-art GIS functionality. This provides the basis for spatial storage and querying at an algebra level. The functionality of this module will be compliant with the Simple Feature Specification for SQL (ISO19125/2). Early implementation of this base functionality will not only be useful for the work at TUDelft but also for all other RGI project that need basic DBMS access. (**Deliverable:** A working spatial module in MonetDB that passes compliance tests for ISO 19125/2 that is available at sourceforge.com)
- **WP2**: During the second half year, CWI will focus on loading, integration, and querying GML/XML data. Several case studies will be undertaken to identify functional, performance and saleability issues that might arise. (**Deliverable:** a scientific publication that compares the performance of MonetDB with other solutions, and gives a detailed analysis of the results).
- **WP3**: During the second year, two research topics will be explored: integration of spatial feature extractors and approximate querying in spatial world. The former capitalizes upon work in peer-projects (MultimediaN) to assess opportunities for stream based updates of a GML/XML store. The latter is focussed on improving our knowledge of spatial query optimizers. The MonetDB technology provides an open architecture to plug-in both generic and domain-specific optimizers. (**Deliverable:** at least 2 scientific publications on the topic, one of which in a journal)

TUDelft:

- **WP4**: TUDelft will start with a problem description phase. Together with the two market parties the postdoc will study how the current practise is at the RWS and Cyclomedia. And will perform an analysis of the problems that the project members have with current DBMS products (What operations are still done outside the DBMS, and why?). At the end of the inventarization phase the researcher will have a good insight in the challenges that the industry partners experience when using current DBMS products) (**Deliverable**: a conference publication that describes the problems and challenges in the current practise together with proposed solutions. (Q1 + Q2).
- **WP5**: In the second phase the problems will be prioritised and an implementation plan for the system that will be developed. In order to ensure end user applicability the business partners will have a strong voice in this process. The deliverable will be a detailed planning and product description for the rest of the project (Q3). In parallel the basic spatial module that has been finished by the CWI team, will be tested with loading and integrating the spatial data from the RWS and Cyclomedia. (**Deliverable**: A detailed planning. The scientific publications will be postponed to the implementation phase)
- **WP6:** In the implementation phase TUDelft will develop functionality that is useful for the industry partners in the project. A deliverable will be a reference implementation that solves the problems such as fined in the requirements phase. (Q4-Q7). A Detailed planning of this phase will be provided in phase two of this project. No later than in Q6 software that is developed should be published in an open source product at sourceforge.org giving the open source community a chance to used and validate the software.
  In this phase of the project there should be DBMS solutions for problems that have not been solved inside a DMBS before. (**Deliverable:** These solutions will be the basis for at least two scientific publications: One with a focus on the problems of RWS/AGI and the other with a focus on Cyclomedia. These papers will focus on the practical applicability of the solutions and will be co-authored by the members of the business partners.)
- **WP7**: At the end of the research the results of the project will be evaluated together with the industry and a workshop will be held where the results of the project will be shared with other parties. (**Deliverable:** Journal publication with the general lessons learnt from the whole project)

Industry Partners:

- The role of the industry partners is described in the planning above, they can participate in all phases of the project.

Workshops:

During the project four workshops will be organized. Two workshops per year:

- The first workshop will focus on problem definition. In this workshop strong participation from both RWS and CycloMedia is essential because at this workshop it is investigated what DBMS functionality is needed the most.
- In the second workshop the basic GIS implementation in MonetDB will be presented. From this moment the DBMS will be available to the RGI projects and the rest of the world evaluation.

- The third workshop will be about the ongoing research. A suitable hot topic will be chosen later.
- The fourth workshop will take place at the end of the project and will aim at transferring knowledge from the project to the world.

Risks Analysis:

- **Risk**: The postdocs do not work well together.
  Measures: At the beginning of the projects the two postdocs will get to know each other by regular visits to each other's institutes. The project leader will supervise the collaboration of the postdocs. The work packages of the postdocs will be split along natural break-lines so that the results of their work are not too much interdependent.
- **Risk**: After the first workshop there is still no agreement on what to implement.
  Measure: The project leader will decide.
- **Risk**: The production of the base spatial module at CWI is delayed. This blocks the progress of the TUDelft postdoc.
  Measure: At this stage only the basic functionality is needed that is also available in other products. The postdoc can proceed using one of these products and migrate later.

Securing knowledge and communication (max. *500 words)*

- *Describe all activities for transferring and embedding the knowledge developed within the project and describe the division of tasks within the project consortium.*
- *If applicable, describe the envisaged approach to marketing the project results*
- *If applicable, describe the supplementary agreements made for intellectual property and the associated restrictions on the disclosure of project results*
- *Relationships: Identify any existing relationships with other projects within Ruimte voor Geo-Informatie, with other Bsik programmes and with other national or international research initiatives*

In the planning paragraph of the project many of the deliverables are aimed towards securing knowledge and communication. Here we sum them up:

- 4 workshops.
- 7 conference contributions.
- 4 scientific journal contributions
- 4 contributions to professional journals

Beside that we have the following plans for communication:

Communication among the team members is organized using regular visits and a common accessible Wiki.

During the whole project there will be publications in relevant scientific journals conferences etc. The reference-architecture will be available for use for other RGI partners. The software products will be included in the distribution of the open source DBMS MonetDB (http://sourcforge.net/projects/monetdb). In this way a long term (beyond the end of the project) international availability and support for the software can be guaranteed.

The total cost of the project is shared between the partners. Their contributions in time and commercial license fee of the data sets made available to the project is used as matching for the cost of the two postdocs. A limited budget is reserved for dissemination.

Relations with other projects:

GeoInfoNed has relationships with the following projects:

- RGI-116: Exploration of innovations in geo-standards for SDI:
  This project aims at the active participation in standardization in geo-information. During the course of our project we will use standards as much as possible, however because of the innovative nature of the project we will probably run out of the boundaries of what has been standardized. Here we can valuable input to the programme that pursues standardization.
- RGI-233: Usable (and well scaled) mobile maps for consumers.
  This project aims at building a client-server application for map-generalization. The server part of this project can use the platform as provided by GeoInfoNed as a starting point. GeoInfoNed can use the problem specifications of RGI-233 as input for the development of generic functionality.
- OSOSS: This project fits perfectly with the Dutch government programme OSSOS that encourages the use of open standards and open source software for the government. Since the plans of GeoInfoNed are completely inline with the goals of OSOSS we plan to contact OSOSS as soon as possible to see where we can help each other.
- PostgreSQL + PostGIS: This open source projects aims at making a stable database platform with functionality comparable to commercial DBMS systems. Since our goal is to go beyond what is possible in current DBMS w

GeoInfoNed creates a working relationship with the following Bsik programs:

- MultimediaN: a large consortium (http://www.multimedian.nl) aimed at improving multimedia document analysis, video content analysis and search, and novel techniques to improve safety. This consortium has strong links into both hardware suppliers and end-user applications. GeoInfoNed has a direct link multimedia database technology and information retrieval projects.
- Bricks: a consortium led by CWI and NWO (see http://www.bsik-bricks.nl) which aims at stimulation of long term fundamental research in computer science. GeoInfoNed has a direct link with the projects aimed at large scale data-mining and personal information systems.

Contribution to RGI objectives *(max. 500 words)*

RGI objective	Contribution to proposed project
More demand-driven	In a demand-driven environment the efficiency of servers becomes much more important. This (seemingly contradiction) is because data now needs to be served how the client wants it at the time the client wants it. This asks for more flexibility at the server.

A freer flow of knowledge	All activities of the GeoInfoNed project are aimed towards openness and dissemination of knowledge.
More cohesion	The integration of various types of data in one DBMS not only brings data together but also the different fields of expertise that come with the data.
More innovation	Because stability is very important for developers of commercial DBMS developers they are conservative by nature and their architecture has not changed much over the years. In the meantime database hardware has changed very much. MonetDB has proven that a complete overhaul of the DBMS architecture can result in incredible performance gains. The GeoInfoNed project will demonstrate this to the Geo Information community.
Greater awareness	With the database at the core of modern organizations a spatially enabled DBMS with efficient operations will ensure that spatial analysis will become core business instead of something very complex that is done at the GIS department.

Appendices

- *Detailed budget in accordance with accompanying spreadsheet*
- *CV of the intended project leader*
- *CVs of the key researchers*
- *Scientific references: publications, posters, prizes, etc.*
- *Social references: projects, applications, partners, etc.*
- *Contribution of project to Senter milestones in accordance with accompanying table*
- *Copies of the quality certificates of all consortium partners*
- *Admission agreement signed by all consortium partners*