Institution:	**University of Oxford**
Unit of Assessment:	**11 Computer Science and Informatics**
Title of case study:	**Semmle: a powerful query language for analysing large data sources (4)**

1. Summary of the impact (indicative maximum 100 words)

Semmle is a successful spin-out company set up by members of the UoA, based on their research on program analysis. Semmle markets an industrial-strength product allowing organisations with large software systems to understand and manage their code bases. This business intelligence platform started to be sold to prominent customers in 2008, including [text removed for publication] NASA. NASA used it to help ensure the safe landing of the Curiosity Mars Rover.

2. Underpinning research (indicative maximum 500 words)

From the late 90s, the Programming Tools team in the UoA researched ways to implement program analyses in a declarative style. Initially the emphasis was on functional programming, and the first major achievement was the invention of an efficient higher-order matching algorithm by Ganesh Sittampalam and Oege de Moor in 1998. This was an example of selecting a pattern language at the optimum point between expressivity (a bit more than 2nd order but not all of 3rd order) and efficiency. Also, the implementation heavily relied on ideas from logic programming [1].
The next major step was the development of a language for implementing optimising compiler transformations based on a combination of rewriting and side conditions phrased in temporal logic, by David Lacey and Oege de Moor in 2001 [2, 3]. A salient feature was the ability to bind free variables while matching the temporal logic formulae. The link to restricted forms of logic programming became even more apparent at this point, and joint work by Sittampalam, Lacey, De Moor and others explored this connection to express particular kinds of recursive analyses as logic programs that query control flow graphs, with publications in 2002 and 2003. Again the key was to strike the right balance between expressivity and efficiency, by not using the full power of a Turing-complete language.

At that time “aspect-oriented programming” was coming into prominence, and it appeared a perfect area of application for the technologies developed in the Programming Tools team. In a landmark paper in 2003, Damien Sereni and Oege de Moor showed that the earlier work with Sittampalam could achieve dramatic speedups in the execution of aspect-oriented programs. This was the beginning of the work on aspect-orientated program optimisation.

In 2004 a new optimising compiler for the most popular aspect-oriented programming language, AspectJ, was implemented by the Programming Tools team, demonstrating that the theoretical speedups could be achieved on an industrial scale [4]. In the meantime a number of other research groups had started using the Datalog programming language, a much restricted logic programming language that is a subset of PROLOG, for the purpose of program analysis. In the UoA, Elnar Hajiyev, together with Mathieu Verbaere and Oege de Moor, provided a Datalog implementation for experimentation using traditional database technology instead of the binary decision diagrams used by others, building on previous work by Thomas Reps at Wisconsin (who also used it for program analysis applications) and others [5].

It became apparent that no-one properly understood the semantics of the event patterns (called “pointcuts”) in AspectJ, and therefore a clear semantics of its pattern language was required. The Programming Tools team addressed this question by translating that pattern language into

Datalog, and giving a simple implementation through that translation [6].

3. References to the research (indicative maximum of six references)

The three asterisked outputs best indicate the quality of the underpinning research.

[1] Oege de Moor, Ganesh Sittampalam: Generic Program Transformation. Advanced Functional Programming 1998: 116-149. DOI: 10.1007/10704973_3

Here we give a new, practical algorithm for a restricted form of higher-order matching, and show how it is used in implementing the well-known fusion optimisations.

***[2] Oege de Moor, David Lacey, Eric Van Wyk: Universal Regular Path Queries. Higher-Order and Symbolic Computation 16(1-2): 15-35 (2003). DOI: 10.1023/A:1023063919574**

Here we present an extension of logic programming to easily express such analyses as patterns on paths in a control flow graph.

[3] David Lacey, Oege de Moor: Imperative Program Transformation by Rewriting. CC 2001: 52-68. DOI: 10.1007/3-540-45306-7_5

It is generally thought that well-known compiler optimisations are hard to implement by rewriting, but here we show that with side conditions written in temporal logic, the formulations are elegant and easy to reason about.

[4] Pavel Avgustinov, Aske Simon Christensen, Laurie J. Hendren, Sascha Kuzins, Jennifer Lhoták, Ondrej Lhoták, Oege de Moor, Damien Sereni, Ganesh Sittampalam, Julian Tibble: Optimising aspectJ. PLDI 2005: 117-128. DOI: 10.1145/1065010.1065026

In this paper, for the first time the overheads of aspects are considered, and we present analyses and optimisations for eliminating those overheads.

***[5] Elnar Hajiyev, Mathieu Verbaere, Oege de Moor: codeQuest: Scalable Source Code Queries with Datalog. ECOOP 2006: 2-27. DOI: 10.1007/11785477_2**

Here we present an implementation of Datalog that executes on top of a standard SQL database, and benchmark its performance on some common source code querying tasks.

***[6] Pavel Avgustinov, Elnar Hajiyev, Neil Ongkingco, Oege de Moor, Damien Sereni, Julian Tibble, Mathieu Verbaere: Semantics of static pointcuts in aspectJ. POPL 2007: 11-23. DOI: 10.1145/1190216.1190221**

The pattern language for pointcuts in AspectJ (the popular aspect-oriented extension of AspectJ) did not have a published semantics: here that void is filled, by a translation to Datalog.

4. Details of the impact (indicative maximum 750 words)

Path to Impact

Semmle was founded in December 2006 by Oege de Moor through ISIS Innovation to create from scratch the novel technology that realises the potential of the research set out in Section 2 in the context of Datalog, widening the scope of application to business intelligence rather than just program analysis. Further research was done in Semmle and six patents were filed by Semmle to protect these advances after the creation of the company.Semmle's first two years were dedicated to developing its technology and products. Its first major customer was signed up in 2008 [text removed for publication].

Direct Economic Impact

[text removed for publication]. Since November 2011, Semmle has been backed by a private investor group in the San Francisco Bay Area [text removed for publication] Many of Semmle's major clients are located in the US [text removed for publication] [A].

Large software engineering organisations struggle with a lack of visibility of how outsourced development teams are performing. To create such visibility it is necessary to analyse many different data sources, including the code itself, version history, the bug database, and test results. Much of this data are graphs and hierarchies, and thus writing the appropriate analysis to get

visibility is hard. The research described here has (through further innovations at Semmle) led to a query language that, for the first time, opens up all the relevant data. Semmle's technology is routinely used in software production and management at numerous clients around the globe and has had significant benefits for them, as described below.

NASA Jet Propulsion Laboratory (JPL) has highly rigorous quality requirements and makes use of all the available leading commercial products in this area to help to ensure that its software is free of errors. Writing custom checkers is key in this environment. Together with Gerard Holzmann at JPL, Semmle has used its query language to implement his well-known "Power of Ten" rules for C programming. Together with Klaus Havelund, Semmle has implemented a new Java coding standard for use at JPL, and Semmle technology was used as one of a suite of static analysis tools to help secure the safe landing of the Curiosity Rover on Mars in August 2012, a $2.5 billion space programme. As Curiosity travelled towards Mars, computer scientists at JPL continuously tested and adjusted its software. In February 2012 a JPL engineer discovered a previously undetected code defect in the critical Entry, Descent and Landing (EDL) software. The defect had no adverse effect on the functioning of the software, but it highlighted the possibility that similar issues could exist in the code with more severe consequences. Any failure in the EDL system could lead to a catastrophic failure of the entire mission. The standard set of rigorous static analysis checks performed on every build of the code could not detect the flaw, and JPL contacted Semmle directly for help. Semmle created a custom analysis that, when run on the code that controlled the spacecraft, quickly identified the known issue and a few related cases, giving the engineers time to fix the code and helping to secure a safe landing [B]. In May 2012 Gerald J. Holzmann, one of the two leaders of NASA JPL's Laboratory for Reliable Software, published a ranked assessment of static source code analysis tools on his website placing Semmle at the top with the comment: '*The tool gives accurate results and, once the database is built for a new application, queries are resolved very fast. Highly recommended.'* [C]

Certipost is an important player in the European market for electronic communication and document processing, used by 85,000 companies and 520,000 individuals. Originally Certipost attempted to improve software architectures with open source tooling, but this was found ineffective because it was too generic: custom analyses were required. In particular Certipost wanted to check that the state of the code was in accordance with the design diagrams; using Semmle's query language, this was achieved with a minimum of effort. In 2012 Certipost published a paper describing the way in which Semmle technology was customised with rules encoding Certipost architecture, thus enabling them to keep code at a high level of quality and in sync with the defined architecture [D]. The head of architecture and analysis at Certipost has commented: 'Certipost can now cross-check code and architecture against each other and detect gaps quickly and easily. The benefit is a reduced cost of keeping these in sync, keeping key quality attributes at desired levels. Semmle has become an indispensable solution for Certipost.'

[text removed for publication]Murex is the 2nd largest independent software vendor in France. It has a large code base, which other products had difficulty processing. Semmle's efficient queries were able to process it without problems, and again custom queries that linked code and other artefacts (in this case XML descriptions of screens) were imperative. Semmle's query language also proved valuable in implementing complex, large-scale refactorings. It is estimated that Semmle has saved them a 30% reduction in developer time on structural improvement and a 10% reduction in corrective maintenance on products that use Semmle. Murex have publically stated the benefits of using Semmle technology; their COO states: "The insight provided by Semmle has been critical to

the success of several improvement projects at Murex. We find the ability to formulate custom analyses a key advantage. Semmle business intelligence is essential for a clean and clear understanding of complex software engineering data." [E]

EMC Corporation is an American multinational corporation that sells data storage products and services used to build web-based computing systems. Its big data division uses Semmle to impose rigorous coding standards, especially in the mission-critical query optimiser. EMC trialled products from vendors that are much longer established, but opted for Semmle because of the greater accuracy, which is enabled by the special query language. The Director of Software Engineering at Greenplum, a subsidiary of EMC, states: 'Semmle has changed my view of software engineering. Semmle ensures a tight adherence to our standard of software excellence. New team members get help in avoiding common beginners' mistakes, and experienced developers can spread their knowledge of good practice.' [F]

5. Sources to corroborate the impact (indicative maximum of 10 references)

[A] Information about Semmle can be corroborated by Oege de Moor, its founder and CEO; information about clients is also corroborated on the Semmle website at http://semmle.com/customers/

[B] http://semmle.com/?case-study=nasa-jet-propulsion-laboratory-jpl
This article from the SEMMLE website confirms the use of Semmle technology to help secure the safe landing of Curiosity on Mars. Also confirmed at http://gigaom.com/2012/08/20/nasa-scrubbed-mars-rover-code-clean-over-and-over/

[C] Corroboration of Gerald J. Holzmann's positive views on Semmle are found on his website at http://spinroot.com/static/ and on Semmle's website at the end of the NASA JPL case study at http://semmle.com/?case-study=nasa-jet-propulsion-laboratory-jpl

[D] De Schutter, K. Automated architectural reviews with Semmle. 28th IEEE International Conference on Software Maintenance (ICSM), 2012. DOI: 10.1109/ICSM.2012.6405320.
A key paper presented by Certipost at ICSM, the premiere international event in software maintenance and evolution, confirming the use of Semmle's custom analyses to avoid the common problem of architectural drift. Corroboration of the benefits to Certipost of Semmle's technology is also contained in a case study on the Semmle website at http://semmle.com/?case-study=certipost

[E] Jean-Pierre Dacher, COO of Murex and former VP of Engineering at SAP Business Objects, confirms the benefits to Murex of Semmle's technology (including the reductions in developer time and corrective maintenance) in a case study on the Semmle website at http://semmle.com/?case-study=murex

[F] Florian Waas, Director of Software Engineering, Greenplum, confirms the benefits to EMC of Semmle's technology in a case study on the Semmle website at http://semmle.com/?case-study=greenplum-a-division-of-emc