1

Introduction

To be effective, data-intensive systems require extensive ongoing customisation to reflect changing user requirements, organisational policies, and the structure and interpretation of the data they hold. Manual customisation is expensive, time-consuming, and error-prone. In large complex systems, the value of the data can be such that exhaustive testing is necessary before any new feature can be added to the existing design. In most cases, precise details of requirements, policies and data will change during the lifetime of the system, forcing a choice between expensive modification and continued operation with an inefficient design.

In 2013, the Networked European Software and Services Initiative (NESSI) identified "Collaborative Service Engineering based on convergence of software and data" as an EU research priority. Information systems are composed of software and data components that must co-evolve as requirements change. In existing development methodologies, software and data engineering are considered as separate concerns.[1] New techniques and tools are required to support the development of effective solutions in the presence of changing requirements, policies, schemas, and data. NESSI also identified "Integration of Big Data Analytics into Business processes" as a research priority, emphasising the importance of data-centric or "Big Data" approaches. This serves only to emphasise the relative value of the data and the need for agility. Big Data approaches involve the imposition of multiple, changing models upon unstructured heterogeneous Linked Data. A single static data model will not suffice, and the manual development of customised code against multiple changing models is unsustainably expensive. Automatic support for customisation, driven by domain models of knowledge and requirements, is an essential component of effective, sustainable Big Data solutions, building on underlying technology from both domains.

[1] A. Cleve, T. Mens, J-L. Hainaut, Data-Intensive System Evolution, IEEE Computer, August 2010.

In software engineering, there are meta-modelling frameworks of the kind that support the Unified Modeling Language (UML), allowing engineers to describe and design features that work for whole classes or families of data models, rather than for a specific instance. There is widespread language support for higher-order programming, in which programs are managed as data. There are mature formal program specification approaches and languages that enable programs to be described mathematically and to be provably correct. We have model-driven, product-line, and generative programming techniques, in which a single set of validated transformations is used to produce or customise many different applications or many different versions of the same application. However, evidence is lacking for the effectiveness of these techniques except in narrow domains.[2]

In data engineering, we have meta-formats such as eXtensible Markup Language (XML), allowing us to describe and design data formats and representations. We have languages such as the Resource Description Framework (RDF) for recording and communicating relationships between different data items; Resource Description Framework Schema (RDFS) for detailing relationships between classes of entities; the Web Ontology Language (OWL) for describing domain knowledge, axioms, and inference rules; and powerful, scalable tools for applying knowledge and rules to large collections of data and metadata. These tools overlap with the expressivity of UML, but in practice, the tractability of code or transformation generation and the ability to reuse data from these syntax-focussed expressions are much weaker than those of native semantic models. More important is perhaps the skills and engineering culture gaps that divide the software and data engineering communities. Common tools that bridge this gap will lead to a deeper shared understanding.

The challenge is to bring these aspects together in a practical, proven methodology, which can be instantiated in software, and which enables the effective, sustainable development of large, complex, and data-intensive systems.

1.1 State of the Art in Engineering Data-Intensive Systems

While the topic of co-evolution between software artefacts and other artefacts produced during software development is an active area of research, its

[2] J. Hutchinson et al. "Model-driven engineering practices in industry," Software Engineering (ICSE), pp. 633,642, 21–28, 2011.

application to data-intensive software systems is not trivial.[3] Although the research focus had been fixed firmly on software interacting with traditional data environments of relational databases[4] and data warehousing,[5] recently, a more technology-independent approach has emerged. Mori and Cleve[6] introduced the notion of data-intensive self-adaptive systems as data-intensive systems able to perform context-dependent data access. They proposed adoption of a framework that supports feature-based data tailoring by means of a filtering design process and a run-time filtering process. Manousis et al.[7] introduced a method for the adaptation of data-intensive ecosystems based on three algorithms that (i) assess the impact of a change, (ii) compute the need of different variants of an ecosystem's components, depending on policy conflicts, and (iii) rewrite the modules to adapt to the change.

Naturally, a prerequisite to assessing impact is the ability to represent the interdependency of the artefacts in a machine-processable manner. Terwilliger et al.[8] stated that "bi-directional mappings" are emerging as a mechanism in the software engineering domain to represent such interdependency. They also identify, characterise, and compare a representative set of tools implementing the approach. Compatible with the concepts, but emerging from the data community, are semantic mappings, where progress has been made in representing and characterising complex mappings through correspondence patterns.[9]

---

[3] A. Serebrenik & T. Mens. Emerging trends in software evolution. In Evolving software systems, pp. 329–332, Berlin: Springer, 2014.

[4] A. Cleve, T. Mens, and J.-L. Hainaut, Data-intensive system evolution, IEEE Computer, vol. 43, no. 8, pp. 110–112, 2010.

[5] A. Abelló, J. Darmont, L. Etcheverry, M. Golfarelli, J. Mazón, F. Naumann, T. Pedersen et al. "Fusion cubes: Towards self-service business intelligence." International Journal of Data Warehousing and Mining (IJDWM) 9, no. 2, pp. 66–88, 2013.

[6] M. Mori, A. Cleve, Towards Highly Adaptive Data-Intensive Systems: A Research Agenda, Advanced Information Systems Engineering Workshops, Lecture Notes in Business Information Processing Volume 148, pp. 386–401, 2013.

[7] P. Manousis, P. Vassiliadis, G. Papastefanatos, Automating the Adaptation of Evolving Data-Intensive Ecosystems, Conceptual Modelling, Lecture Notes in Computer Science Volume 8217, pp. 182–196, 2013.

[8] J. F. Terwilliger, A. Cleve, C. A. Curino, How Clean Is Your Sandbox?, Theory and Practice of Model Transformations, Lecture Notes in Computer Science Volume 7307, pp. 1–23, 2012.

[9] J. Keeney, A. Boran, I. Bedini, C. Matheus and P. Patel-Schneider, "Approaches to Relating and Integrating Semantic Data from Heterogeneous Sources." In Proc. 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Vol 01, pp. 170–177. IEEE Computer Society, 2011.

Empirical studies and research that help to motivate the need for strongly integrated system co-evolution are also emerging. Goeminne et al.[10] reported on early results obtained in the empirical analysis of the co-evolution between code-related and database-related activities of contributors in a large open source data-intensive system. Their study investigated questions such as: what is the effect of introducing a new database technology? And how do developers divide their effort between the activity types involved in evolving a data-intensive system? Papastefanatos et al.[11] proposed a set of graph-theoretic metrics for the prediction of impact of schema evolution upon ETL software and evaluated them over seven systems. Meurice and Cleve,[12] in a short study, described the type of schema evolution that emerged in four systems over a period of months and the utility of having a tool to aid the analysis. Sen and Gotlieb[13] proposed a methodology for testing data-intensive systems and present results achieved when applied to a case study in the Norwegian Customs and Excise governmental department.

1.1.1 The Challenge

There is a body of research studying data-intensive systems, from a unified point of view, but the focus to date has been largely on relational data models. These are, of course, important for current enterprise systems. However, the Web is currently undergoing a data revolution, where machine-to-machine communication will eventually dominate over human-centric, document-oriented Web traffic. A key driver of this data revolution is graph-based data, whether in the form of the Facebook Graph API[14] for searching their social graph, Google, Bing, Yandex and Yahoo's schema.org for annotating Web pages with graph-based metadata or the W3C's Linked Open Data (LOD)

[10]M. Goeminne, A. Decan, T. Mens, (2014, February). Co-evolving code-related and database-related changes in a data-intensive software system. In Proceedings of the IEEE CSMR-WCRE 2014 Software Evolution Week.

[11]G. Papastefanatos, P. Vassiliadis, A. Simitsis, Y. Vassiliou, Metrics for the Prediction of Evolution Impact in ETL Ecosystems: A Case Study, Journal on Data Semantics, Volume 1, Issue 2, pp. 75–97, August 2012.

[12]L. Meurice & A. Cleve, DAHLIA: A Visual Analyzer of Database Schema Evolution, CSMR-WCRE 2014, Belgium, 2014.

[13]S. Sen and A. Gotlieb, Testing a Data-intensive System with Generated Data Interactions: The Norwegian Customs and Excise Case Study, 25th International Conference on Advanced Information Systems Engineering (CAISE'13) (2013).

[14]https://developers.facebook.com/docs/graph-api/

community[15] that builds on over a decade of semantic Web research. For the next generation of Web-scale data-intensive systems, it is not enough to transfer legacy data models to the cloud. Instead, the research on controlled co-evolution of software and data must be extended to deal natively with Linked Data-based systems.

Many of the techniques developed for traditional data-intensive systems, such as data transformation generation, are still relevant, but Linked Data versions must be developed. The richer models of semantic, RDF-based methods offer new opportunities: for leveraging domain knowledge expressed as ontologies; applying semantic mapping techniques for correspondence classification to schema evolution evaluation (to drive controlled transformations for programs, queries, and data); and modelling the software and data life cycles in a machine-computable way, enabling heterogeneous tools to collaborate in combined software and data engineering tool chains.

1.2 State of the Art in Semantics-Driven Software Engineering

Model-driven software engineering is the automatic production of software artefacts from abstract models of structure and functionality. This approach can reduce the costs of development and maintenance and increase the quality and reliability of the software produced. It has been adopted for the development of control and embedded systems,[16] for aspects of data warehousing,[17] and for service implementations.[18] It has yet to achieve any widespread adoption outside these domains. Multiple reasons are suggested by Den Haan,[19] but the two most common explanations are a lack of adequate

[15] http://www.w3.org/standards/semanticweb/data

[16] D. Hästbacka, T. Vepsäläinen, S. Kuikka, Model-driven development of industrial process control applications, Journal of Systems and Software, Volume 84, Issue 7, pp. 1100–1113, July 2011.

[17] J. Mazón, J. Trujillo, M. Serrano, and M. Piattini. "Applying MDA to the development of data warehouses." In Proceedings of the 8th ACM international workshop on Data warehousing and OLAP, pp. 57–66. ACM, 2005.

[18] J. Bezivin, S. Hammoudi, D. Lopes, and F. Jouault. "Applying MDA approach for web service platform." In Enterprise Distributed Object Computing Conference, 2004. EDOC 2004. pp. 58–70. IEEE, 2004.

[19] J. Den Haan, "8 Reasons Why Model-Driven Approaches (will) Fail". http://www.infoq.com/articles/8reasons-why-MDE-fails, July 2008.

tool support[20] and, as a consequence, a lack of any proven, empirically tested methodology.

Existing tools are focussed on the production of structural, static components of an implementation. Beyond a handful of tightly constrained domains, these tools lack any means to model and generate anything beyond the most basic aspects of functionality.

Technology platforms are available to support more general model transformation and code production. Many of these have been implemented in the widely used Eclipse environment and address the Object Management Group's (OMG) Model-Driven Architecture (MDA) proposal,[21] with tools for domain-specific modelling,[22] developing model transformations,[23] and performing model edits and manipulations.[24]

The Atlas Transformation Language, in particular, is based on the Query View Transformation proposal[25] for transformation languages and acts on models written in UML: the de facto industry standard for software systems modelling. Techniques have been developed that support genericity and bi-directional transformation,[26] with the aim of facilitating round-trip engineering and iterative development. Specialised tools, such as Stratego,[27] have been developed for program transformation or meta-programming.

---

[20] J. Whittle, J. Hutchinson, M. Rouncefield, B. Håkan, and R. Heldal. "Industrial Adoption of Model-Driven Engineering: Are the Tools Really the Problem?" In Model-Driven Engineering Languages and Systems, pp. 1–17. Springer, 2013.

[21] A. Kleppe, J. Warmer, W. Bast, "M.D.A. Explained. The model driven architecture: practice and promise", 2003.

[22] F. Jouault, J. Bézivin, and I. Kurtev, "TCS: a DSL for the Specification of Textual Concrete Syntaxes in Model Engineering," in Procs of the 5th Int. Conf. on Generative programming and Component Engineering (GPCE '06). New York, NY, USA: ACM, pp. 249–254, 2006.

[23] F. Jouault, F. Allilaire, J. Bézivin, I. Kurtev, ATL: A model transformation tool, Science of Computer Programming

[24] M. Del Fabro, J. Bézivin, and P. Valduriez. "Weaving Models with the Eclipse AMW plugin." In Eclipse Modelling Symposium, Eclipse Summit Europe (2006).

[25] MG, Meta Object Facility (MOF) 2.0 Query/View/Transformation Specification, OMG Document formal/2011-01-01, Object Management Group, http://www.omg.org/spec/QVT/1.1/ (2011).

[26] J. Cuadrado, E. Guerra, and J. De Lara. "Generic model transformations: write once, reuse everywhere." In Theory and Practice of Model Transformations, pp. 62–77, Springer Berlin Heidelberg, 2011.

[27] E. Visser, Program transformation with Stratego/XT, in: Domain-Specific Program Generation, Vol. 3016 of Lecture Notes in Computer Science, pp. 216–238, Springer Berlin Heidelberg, 2004.

There has been work on mappings between the ISO/IEC 11179 metadata registry standard and description logics, such as OWL,[28] but this has focussed purely on the representation of modelling constructs, with no consideration of the implications for software and data engineering. Similarly, within the OMG, efforts have focussed on how to enable the use of UML notation and tools for ontology modelling.[29] There has been related work on representing systems specifications as ontologies for project planning that draws on the OMG MDA specification as inspiration.[30]

The most significant effort to date on the incorporation of semantic models into software engineering has been the FP7 MOST project (2007–2011), which investigated the utilisation of ontologies in an MDA approach.[31] Their work developed new techniques for applying semantic reasoners to MDA tasks,[32] such as model checking, specification validation, or supporting domain specific languages (DSLs) with strong semantics. Much effort was focussed on model translation or bridging[33] between non-mainstream UML variants such as grUML and OWL ontologies. In a 2013 update,[34] one of the project's principal investigators laid out a vision for Ontology-Driven Software Engineering that targets 2030 as the year when this technology will be mature. This timescale indicates the difficulty of building formal ontologies into the heart of software engineering. It also distinguishes this work from the approach of ALIGNED, which is based on a more lightweight Linked Data methodology that aims to enable reuse of rich dataset and metadata descriptions by software engineering tools while supporting co-evolution

[28] C. Tao, G. Jiang, W. Wei, H. R. Solbrig, and C. G. Chute. "Towards semantic-web based representation and harmonization of standard meta-data models for clinical studies." AMIA Summits on Translational Science Proceedings: 59 (2011).

[29] S. Brockmans, R. M. Colomb, P. Haase, E. F. Kendall, E. K. Wallace, C. Welty, G. Tong Xie. A Model Driven Approach for Building OWL DL and OWL Full Ontologies, ISWC 2006.

[30] M. Líška and P. Navrat, An Approach to Project Planning Employing Software and Systems Engineering Meta Model Represented by an Ontology, ComSIS Vol.v7, No. 4, December 2010.

[31] http://www.slideshare.net/malgorzatasiwiec/ontologies-and-software-technologies-the-most-project.

[32] http://www.slideshare.net/fparreiras/filling-the-gap-between-semantic-web-owl-ontology-technology-andmodel-driven-engineering-mde-mdsd-mda.

[33] T. Walter, Bridging Technological Spaces: Towards the Combination of Model-Driven Engineering and Ontology Technologies, PhD thesis, Universite Koblenz-Landau, 2011.

[34] U. Assmann, Current Trends and Perspectives in Ontology-Driven Software Development, August 2013, available at http://www.computational-logic.org/content/events/iccl-ss-2013/download/assmann-1-odsd.pdf.

of software and data assets. In 2012, Katasonov[35] pointed the way forward, "beyond model checking and transformations", with a call to apply semantics in software engineering for its known capabilities in describing software and data assets, as well as semantic search and multi-layered modelling of systems.

1.2.1 The Challenge

There is a large body of research on model-driven engineering (MDE), and, in principle, its benefits are clear, especially for evolvable systems. Despite this and the high-profile OMG MDA initiative of the early 2000s, it has not succeeded in proliferating to the mainstream of software engineering practice other than in embedded systems and certain niches. Modern data-intensive systems are characterised by the need to meet changing application requirements and to integrate multiple data sources whose ownership may lie outside the authority of the application developers. The goal of the ALIGNED project was to change this by collecting quantitative evidence of the benefits of deploying model-driven technology in enterprise information processing systems. The basis of this was aggregating formal system specifications for both data and software, based on a common set of metamodels or vocabularies.

There is already evidence that ontologies or semantic models can provide benefit as input domain models for model-driven development. Despite this, semantic data engineering is a marginal activity at the periphery of software engineering. There is an opportunity to create a more holistic view of the data-intensive system engineering process. By modelling design intents, life cycles, and inter-life cycle communication, it was possible to better integrate the tools and methods used in the software and data engineering processes, in order to enable loosely coupled co-evolution of systems and external Web data resources.

1.3 State of the Art in Data Quality Engineering

Data quality engineering is an issue that exists independently of data representation and technology and arises wherever data are stored for incorporation into business processes. However, in general, the older and more

[35] A. Katasonov, Ontology-driven software engineering: Beyond model checking and transformations, International Journal of Semantic Computing, Vol. 6 (2012) No. 2, pp. 205–242, 2012.

established a language and technology, the more mature the tools, standards, and processes are for dealing with data quality engineering issues. For example, where XML is concerned, Schematron[36] is an ISO standard for validation and quality control of XML documents based on XPath and XSLT. Similarly, in database research, there are related approaches to formulate common integrity constraints[37] using First Order Logic (FOL). The work of Fan,[38] for example, uses FOL to describe data dependencies for quality assessment and suggests repairing strategies. The development of similar mechanisms for RDF is of crucial importance to provide solutions to allow the use of RDF in settings that require either high-quality data or at least an accurate assessment of its quality.

Several approaches for assessing the quality of Linked Data have been proposed, which can be broadly classified into (i) automated;[39] (ii) semi-automated;[40] and (iii) manual[41] methodologies. These approaches introduce systematic methodologies for assessing the quality of an RDF dataset at the process level. Additionally, there have been efforts to assess the quality of large-scale Web data,[42] which included the analysis of 14.1 billion HTML tables from Google's general-purpose Web crawl in order to retrieve tables with high-quality relations. Similarly, Hogan et al.[43] assessed the quality of published RDF data. This study described the errors characteristically associated with publishing RDF data, catalogued the available techniques to improve the quality of structured data on the Web, and analysed each technique's effectiveness. In a recent study, 4 million RDF/XML documents were analysed, which provided insights into the level of conformance these

[36] http://www.schematron.com/

[37] A. Deutsch. Fol modelling of integrity constraints (dependencies). In L. LIU and M. ÖZSU, editors, Encyclopedia of Database Systems, pp. 1155–1161, Springer US, 2009.

[38] W. Fan. Dependencies revisited for improving data quality. In Proceedings of the Twenty-seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, ACM, pp. 159–170, New York, NY, USA, 2008.

[39] C. Guéret, P. T. Groth, C. Stadler, and J. Lehmann. Assessing linked data mappings using network measures. In Proceedings of the 9th Extended Semantic Web Conference, volume 7295 of LNCS, pp. 87–102. Springer, 2012.

[40] A. Flemming. Quality characteristics of linked data publishing datasources. MSc thesis, Humboldt-Universität Berlin, 2010.

[41] C. Bizer and R. Cyganiak. Quality-driven information filtering using the WIQA policy framework. Web Semantics, 7(1), pp. 1–10, January 2009.

[42] M. J. Cafarella, A. Y. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. Webtables: exploring the power of tables on the web, PVLDB, 1(1), pp. 538–549, 2008.

[43] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. In LDOW, 2010.

documents had with the Linked Data guidelines. This effort assessed a vast amount of Web and RDF/XML data; however, most of the analysis was performed automatically, thereby overlooking the problems arising due to contextual discrepancies. In earlier work, similar ideas were used for describing knowledge base evolution.[44]

The approach described in Fürber and Hepp[45] advocates the use of SPARQL and SPARQL Inferencing Notation (SPIN) for RDF data quality assessment. However, their approach requires a domain expert for the instantiation of test case patterns. SPIN[46] is a W3C submission aimed at representing rules and constraints on Semantic Web models. SPIN also allows users to define SPARQL functions and reuse SPARQL queries. In a similar way, Fürber et al. also defined a set of generic SPARQL queries to identify missing or illegal literal values and datatypes and functional dependency violations. Another related approach is the Pellet Integrity Constraint Validator (ICV).[47] Pellet ICV translates OWL integrity constraints into SPARQL queries. The execution of those SPARQL queries identifies violations. An implication of the integrity constraint semantics of Pellet ICV is that a partial unique names assumption (all resources are considered to be different unless equality is explicitly stated) and a closed world assumption are adopted. qSKOS defines rules to detect potential quality problems in datasets using the Simple Knowledge Organisation System (SKOS) schema. The rules are based on existing thesaurus construction guidelines and are evaluated using SPARQL queries and graph algorithms (e.g., to find weakly connected components). Finally, Lausen et al.[48] suggested extensions to RDF by constraints akin to RDBMS in order to validate data using SPARQL as a constraint language. This is achieved by providing an RDF view on top of the data.

---

[44]C. Rieß, N. Heino, S. Tramp, and S. Auer. EvoPat – Pattern-Based Evolution and Refactoring of RDF Knowledge Bases. In Proceedings of the 9th International Semantic Web Conference (ISWC2010), LNCS, Berlin/Heidelberg, Springer 2010.

[45]C. Fürber and M. Hepp. Using SPARQL and SPIN for data quality management on the semantic web. In W. Abramowicz and R. Tolksdorf, editors, BIS, volume 47 of Lecture Notes in Business Information Processing, pp. 35–46, Springer, 2010.

[46]H. Knublauch, J. A. Hendler, and K. Idehen. SPIN – overview and motivation. W3C Member Submission, February 2011.

[47]E. Sirin and J. Tao. Towards integrity constraints in OWL. In Proceedings of the Workshop on OWL: Experiences and Directions, OWLED, 2009.

[48]G. Lausen, M. Meier, and M. Schmidt. SPARQLing constraints for RDF. In Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology, EDBT '08, ACM, pp. 499–509, New York, NY, USA, 2008.

While there has been considerable research into quality assessment of Linked Data sets, work that attempts to incorporate such efforts into quality engineering frameworks, which operate to improve data quality over time, is only starting to emerge. Feeney et al.[49] described a semi-automated methodology, framework, and process, which integrate RDF quality assessment mechanisms with human workflows for achieving quality control of published RDF datasets.

1.3.1 The Challenge

The challenge that ALIGNED faced in data quality engineering was twofold. First, the data quality engineering processes that the partners developed for Linked Data required further development, validation, and standardisation. Secondly, mechanisms were required to allow quality control actions of software and data teams, which have generally been developed in isolation, to be aligned and synchronised. For example, if a customer bug report arrives, then it can often be solved by modifications in either the applications or the data. How is this responsibility allocated in diverse teams and what solution will have the best outcome in terms of both the short- and long-term agility and integrity of the combined system?

When data quality is vital, the ultimate resource to deploy is human expertise. In some cases, it may be necessary to deploy human experts to annotate and interpret datasets in order to elevate the raw data to useful information or knowledge for the planned application tasks. However, this is very expensive in terms of both time and the limited resource of domain expertise. Fully automated solutions are popular in research applications, but in enterprise, the deployment of human talent dominates. This is because of the persistent gulf in quality between human-curated content and automated approaches. Thus, the challenge for pragmatic systems is to define semi-automated methods and tools that involve human expert curators in the loop while minimising their workload. By partitioning curation tasks into different levels of required expertise, it is possible to lower the expertise required for participation in the data processing pipeline and thus broaden the base of contributors, hence lowering costs and increasing the productivity of the highest-value experts. Curation workflow tools that provide this functionality

[49] K. Feeney, D. O'Sullivan, W., Tai, R. Brennan, Improving curated web-data quality with structured harvesting and assessment (2014), International Journal on Semantic Web and Information Systems.

based on an explicit data life cycle model will result in higher-quality systems at lower cost.

One of the attractions of Linked Data, from an enterprise point of view, is the widespread availability of compatible datasets with which to enrich or annotate an application-specific dataset. However, in practice, this is often seen as an advantage that is still to be realised, since the quality of datasets published on the Web varies widely and it is only recently that mature Linked Data quality frameworks have appeared. Importing low-quality datasets often results in a large clean-up exercise for the application owners. Given that system integrity depends directly on the quality of data input, there is an opportunity to control dataset integrity by limiting updates to datasets based on a strong, semantic specification of the system, the application and schema needs, and design intents. A repository integrity gateway could utilise both data quality frameworks and the system specification to limit the data input, referring offending data to human administrator-based intervention or to other automated checks.

Just as unit testing has entered the mainstream of software development, it is possible to create automated data testing based on rich models of domains, application data needs and design intents and to integrate these into semi-automated processes, which maximise the utilisation of new technologies without dispensing with the ability to use human expertise to provide the highest-quality data. Developing and validating processes that successfully integrate these processes was the challenge tackled by ALIGNED.

1.4 About ALIGNED

ALIGNED is an EU research project, which ran from February 2015 to January 2018. It brought together world-class researchers, representing stakeholders from across the value-chain. It combined model-driven software engineering (Oxford are leading the development of the next generation of UK National Health Service systems), Linked Data quality (Leipzig and Trinity College have published foundational papers) with innovative enterprises (Wolters Kluwer has pioneered the use of Linked Data in complex mission critical systems; the Semantic Web Company (SWC) leads the world in enterprise Linked Data), and expert-driven data curation (Oxford Anthropology and Poznań) to work on high-impact use cases such as DBpedia (Leipzig

are co-creators). The project's ambition was to develop the foundations for the next generation of Big Data systems by enabling model-driven creation of Linked Data applications that can effectively deal with the dynamism, complexity, scale, and data quality challenges (e.g., inconsistency and incompleteness) of Web data while retaining the reliability, security, and robustness that come with model-driven software engineering.

The objective of the ALIGNED project was to align semantics-based model-driven software engineering with full life cycle Linked Data engineering to produce powerful and flexible service engineering systems and enable rapid development cycles based on reuse and extension of heterogeneous data sources. This approach supports an aligned engineering process spanning the full service life cycle, based on rich, semantic Linked Data representations, which enable expressive models to be specified for open extensible systems in such a way that flexibility and reusability are prioritised. This will facilitate a step change in the development[50] of Web-scale data-intensive systems. Successfully attaining this objective requires innovations in three distinct technical areas:

- Model-driven software engineering is a maturing research field with well-developed tools and methods like UML, XML, and DSL creation, code, and transformation generation tools like Stratego/Spoofax.[51] The ALIGNED project evolved this research with more expressive and shareable data models based on the modern Web of data.
- Enterprise Linked Data-based systems are starting to appear,[52] and while Linked Data quality engineering processes have started to emerge,[53] they suffer from inadequate tool support. Most Linked Data life cycle management tools also suffer from being oriented towards knowledge engineers, specialising in semantics, rather than the domain experts or software engineers that build and administer enterprise data-intensive systems. ALIGNED addressed this shortcoming by developing, testing, and validating collaborative Linked Data engineering tools and integrating them into user-friendly data curation services and platforms.

[50] http://www.uml.org/ & http://www.w3.org/XML/

[51] http://strategoxt.org/view/Spoofax/WebHome

[52] C. Dirschl, K. Eck, and J. Lehmann, "Supporting the Data Lifecycle at a Global Publisher using the Linked Data Stack", ERCIM News, 96, January 2014.

[53] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, "Quality assessment methodologies for linked open data", under review, Semantic Web Journal, IOS Press.

- Linked Data schemas, expressed in standardised languages such as RDFS[54] and OWL,[55] enable self-describing data structures with rich semantics included within the data itself. Aspects of program logic previously encapsulated in software are now embedded in data models, meaning that the software engineering life cycle of data-intensive systems needs to be aligned with the data engineering life cycle. For example, changes to data schemas may require updates to the software that consumes it, and vice versa. ALIGNED addressed this challenge by identifying common phases and signalling between the parallel processes and tools to support alignment at higher levels. This supports both integrated, model-driven unified processes and loosely coupled, co-evolving systems through the specification of common vocabularies and domain-specific metamodels.

ALIGNED leveraged Linked Data as the common technical platform to support integration at three levels: first, by applying semantics and Linked Data to model-driven software engineering to develop rich domain and application-specific specification models; second, as a means to integrate tools for combined software and data engineering; and third, as the basis for exemplar data-intensive systems that combine software and data to manage, publish, process, and consume data.

NESSI has identified "Collaborative Service Engineering based on the convergence of software and data" and "Integration of Big Data Analytics into Business processes" as EU research priorities.[56] This is a response to the parallel trends which see increasingly complex and dynamic service-delivery collaborations alongside the ongoing explosive growth of data available via the Web. The increasing prevalence of rich and flexible standardised semantic languages[57] has created opportunities for service providers to add value to their services with readily available machine-processable knowledge.[58] To take advantage of these opportunities, service and software engineering organisations must integrate data engineering and service engineering processes.

---

[54] http://www.w3.org/RDF/ & http://www.w3.org/TR/rdf-schema/

[55] http://www.w3.org/TR/owl2-overview/

[56] Strategic Research and Innovation Agenda Version 2.0, NESSI Position Paper, April 2013.

[57] C. Bizer, K. Eckert, R. Meusel, H. Mühleisen, M. Schuhmacher, and J. Völker: Deployment of RDFa, Microdata, and Microformats on the Web – A Quantitative Analysis In: 12th International Semantic Web Conference, 21–25 October 2013.

[58] P. Hitzler, K. Janowicz, Linked Data, Big Data, and the 4th Paradigm, Semantic Web Journal, IOS Press, 2013.

1.5 ALIGNED Partners

1.5.1 Trinity College Dublin

Trinity College Dublin is Ireland's leading university. TCD, founded in 1592, hosts over 15,500 students. It enjoys an esteemed reputation in research and innovation with an outstanding record of publications in high-impact journals and a track record in winning research funding. Its research impact is currently ranked 44th in the world by the Times Higher Education Ranking of World Universities and 9th in Europe by the 2013 Leiden University Ranking of World Universities' research performance.

1.5.2 Oxford University – Department of Computer Science

The Department of Computer Science, ranked first in Europe in the Shanghai tables, has particular strengths in software engineering, programming languages, and information systems. The Software Engineering Group works across all three areas and has a strong track record of interdisciplinary collaboration in medical and scientific research, humanities, and social sciences. It has also a strong track record of effective engagement with industry, delivering a substantial programme of advanced education aimed at full-time professionals: designers, developers, managers, and users.

1.5.3 Oxford University – School of Anthropology and Museum Ethnography

The School of Anthropology and Museum Ethnography is one of the oldest and most distinguished anthropology departments in the world. It is also one of the broadest, comprising five units that cover a wide range of subfields of anthropology (social and cultural, cognitive and evolutionary, visual and material, medical and biological) as well as a range of specialised foci (e.g., migration, science and technology) with long-established field projects all around the globe. Of particular importance for ALIGNED, it is home to the Institute of Cognitive and Evolutionary Anthropology, which employs staff with expertise in database construction and analysis.

1.5.4 University of Leipzig – Agile Knowledge Engineering and Semantic Web (AKSW)

The Institute for Applied Computer Science (InfAI, http://infai.org) at Universität Leipzig hosts world-class research groups in service and Web science.

The approximately 40 researchers of the Agile Knowledge Engineering and Semantic Web research group (http://aksw.org) at InfAI are establishing theoretical results and scalable implementations for realising the Semantic Data Web. Particular emphasis is given to areas such as ontology creation and manipulation, knowledge extraction, ontology learning and information, and data integration on the Linked Data Web. The scientific publications of the group, founded in 2006, have already attracted more than 6,000 citations (according to Google Scholar).

1.5.5 Semantic Web Company

SWC is an SME, based in Vienna, Austria, founded in 2001, which offers ICT consulting services and solutions in semantic information management. This includes data and metadata management, knowledge and information management systems, LOD, enterprise search, and social software. SWC is the vendor of the PoolParty Semantic Suite (http://poolparty.biz) for enterprise-ready solutions in taxonomy management and data integration. SWC's work is always based on open semantic Web standards to ensure interoperability and sustainability for solutions.

1.5.6 Wolters Kluwer Germany

Wolters Kluwer Germany is an information services company specialising in the legal, business, and tax sectors. Wolters Kluwer provides pertinent information to professionals in the form of literature, software, and services. Headquartered in Cologne, it has over 1,200 employees located at over 20 offices throughout Germany, conducting business on the German market for over 25 years. Wolters Kluwer Germany is part of the leading international information services company, Wolters Kluwer n.v., located in Alphen aan den Rijn (the Netherlands). The core market segments, targeting an audience of professional users, are legal, business, tax, accounting, corporate and finance services, and healthcare.

1.5.7 Adam Mickiewicz University in Poznań

Adam Mickiewicz University in Poznań is the major academic institution in Poznań and one of the top Polish universities. Its reputation is founded on tradition, the outstanding achievements of the faculty, and the attractive curriculum offered to students. It is a centre of academic excellence, where research and teaching are mutually sustaining, and where the context within

which research is conducted and knowledge is sought and applied is international as much as regional and national. The University was founded in 1919 and its current student population is nearly 49,000. The University currently employs nearly 3,000 teaching staff, including 264 tenured professors, 439 associate professors, and 1,617 adjunct professors and senior lecturers.

1.5.8 Wolters Kluwer Poland

Wolters Kluwer Poland the largest publisher of legal and business information in Poland. It provides a large database of legal and business information under the IPG brand. Wolters Kluwer Poland is part of the leading international information services company, Wolters Kluwer n.v., located in Alphen aan den Rijn (the Netherlands).

1.6 Structure

The remainder of the book is organised as follows. Chapter 2, Use Cases, briefly describes the five use cases undertaken in the book. It focusses on the data engineering and software engineering challenges, where they are the same and where they differ across the use cases. Chapter 3, Methodology, describes a general methodology for understanding Big Data systems, their requirements, the different families of modelling approaches that are suitable for different systems, and the integration of software and data engineering life cycles by way of signalling points and common vocabularies. Chapter 4, Vocabularies and Ontologies, describes the use of layered common taxonomies, vocabularies, and ontologies as a basis for semantic integration. These include foundational schemas such as RDF, RDFS, and OWL; common widely used standards such as PROV and SKOS; new general-purpose ontologies to describe validation errors and dataset identities such as RVO and DataID; and high-level custom ontologies to describe processes (DLO and SLO). Chapter 5, Tools, describes the software tools used to solve the problems of the use cases, which include RDFUnit, DataID, the Model Catalogue, Semantic Booster, the PoolParty Semantic Suite, and the Dacura semantic curation platform. It focusses on describing the vocabularies and APIs supported by each tool with a little bit on implementation for each.

Chapter 6, Integrated Systems, describes the integrated systems that were developed to solve the problems of the use cases introduced in Chapter 2.

It is split into five parts:

- Wolters Kluwer – Re-engineering a complex relational database application: In every enterprise environment, relational databases are used for a long time to process critical data. It is a common situation that the database schema has heavily evolved over time and no one in the company understands the impact of any change in its entirety anymore. Therefore, companies continue to use these databases without touching them anymore, reducing its overall value over time. Sooner or later, a complete re-engineering or even complete new development is required, which means a significant investment and a high risk of failure. In this presentation, we will show that it is possible to reduce this risk by using semantic technologies when replacing the old application and which also better prepares the company for any re-engineering effort in the future.
- Seshat – collecting and curating high-value datasets with the Dacura platform: This section uses the Seshat project as a case study – a huge distributed effort by social scientists to compile an authoritative databank describing the evolution of all human societies that have existed since 10,000 BCE. We show how the system uses semantic models both to provide strong data consistency assurances and to generate user interfaces for crowd-sourcing and human expert approval. Although this use case is an academic endeavour, the technology is entirely agnostic to the application and can be applied in any scenario where an organisation wishes to collect and curate high-quality datasets.
- Managing data for the NHS: This section examines the ALIGNED Data Catalogue system: a set of tools for automating aspects of data management at scale. At the heart of the system is the metadata catalogue, a tool for capturing and linking key information about data: information that can be used to determine, automatically, how data are to be processed, transformed, and accessed. Other tools support the processes of metadata capture and curation, as well as system configuration and generation. We explore the application of the Data Catalogue system to the management of health data in the United Kingdom. The Oxford ALIGNED partners have deployed the metadata catalogue and other tools in support of several, large health data projects in collaboration with the NHS. One of these, the 100,000 Genomes Project, required the coordination of data specifications, form designs, database schemas, and messages, for a wide range of diseases, across 70 hospitals.

- Integrating semantic datasets into Enterprise Information Systems with PoolParty: The Linked Data movement has seen increasingly large semantic datasets published on the Web, as part of the web of data. This creates opportunities for integrating public sources of data with enterprise information sources to create enriched high-quality semantic knowledge bases. ALIGNED is developing tools and processes to integrate with PoolParty, SWC's semantic technology suite. PoolParty Thesaurus Server is a Thesaurus and Taxonomy Management Tool to build and maintain information architectures. In this section, we showcase how we use SHACL and the RDFUnit test framework as a basis for the import assistant to run automatically and manually generated test cases for validating data consistency constraints.
- Data Validation at DBpedia: Data validation is a crucial part of data integration – integrated data must meet a minimum validation criterion before it can be considered integrated. Reducing the manual time and effort required to validate data is a critical enabler of dealing with the volume and velocity of Big Data. In this section, we show how DBpedia has used ALIGNED tools including RDFunit to develop a high-quality curated dataset offering.

Finally, Chapter 7, Evaluation, describes a suite of evaluation techniques and measures focussed on agility, productivity, and quality in big-data systems and presents an ontology in which the various types of measures are related to one another and an abstract framework for evaluating such systems.