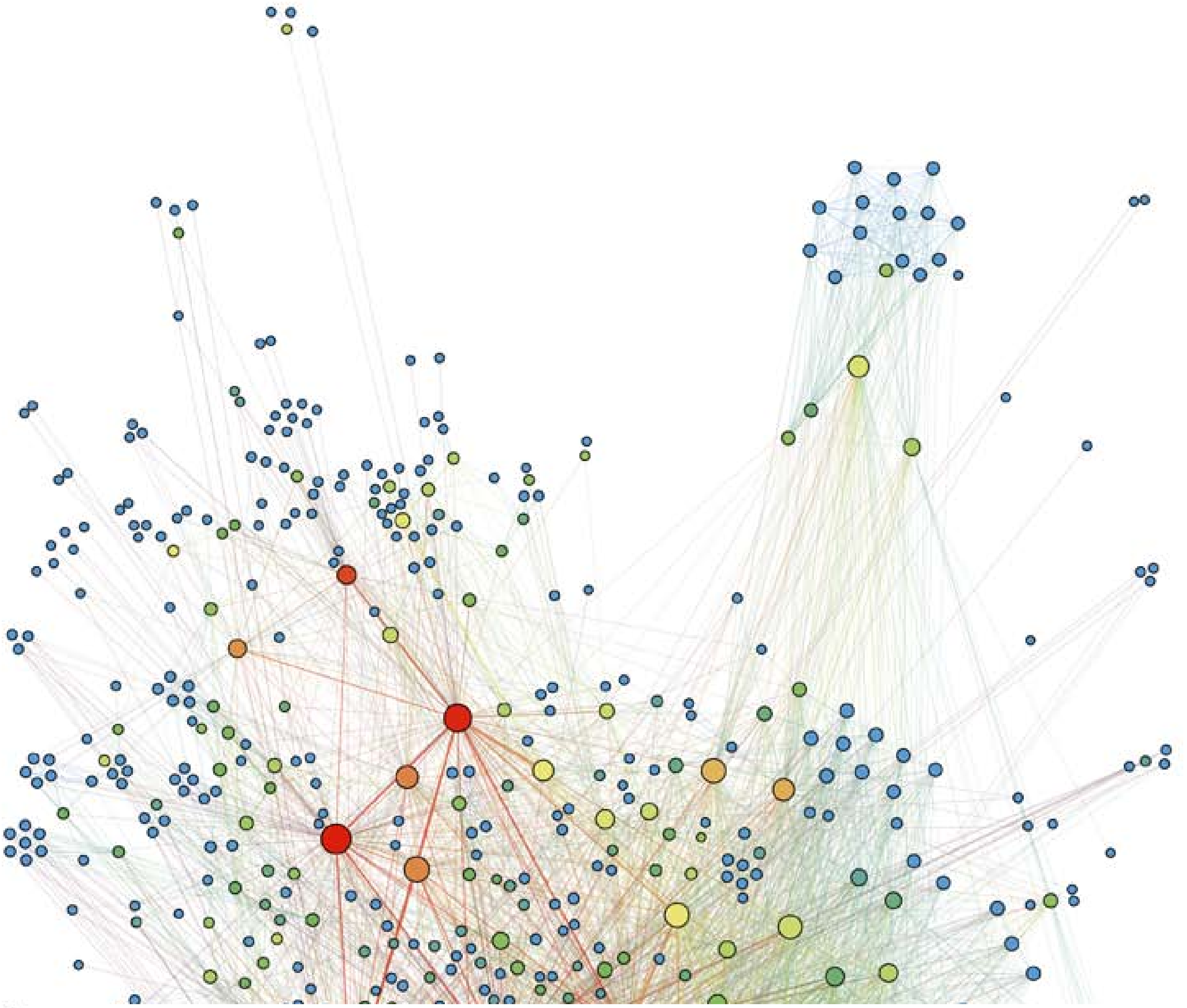


Engineering Agile Big Data Systems

March 20th 2017, 10:00 – 14:00
Hannover Exhibition Grounds, Building CC, Room 3A

ALIGNED
Software and Data Engineering

CeBIT

ABOUT THIS WORKSHOP

In this workshop we will showcase how the latest breakthroughs in computer science research can help companies to deal with the practical challenges of big data, so that they can more efficiently and agilely integrate it into their core business processes. We will present 5 case studies from a range of different sectors: from managing hundreds of millions of clinical-trial results for the NHS in the UK; via providing intelligent productivity enhancing services to legal professionals, based on drawing information from millions of complex legal documents; to one of the world's most ambitious attempts to crowd-source high-quality historical data. In each case we will show how tools developed in the ALIGNED project are enabling large communities of users to leverage the value of their data.

ABOUT THE ALIGNED PROJECT

The ALIGNED project is a European Horizon 2020 research project which has brought together academic researchers and industry partners to develop some of the most promising results in computer science research into practical tools, that can be applied to solve real world problems and help organisations to leverage the great volume, variety and velocity of data that is available internally and through the web. In this workshop, we will demonstrate how data quality analytic tools can be used to apply automated quality control and error correction to promote the veracity of data; how semantic modelling can be used to extract business intelligence from complex, interconnected and fast-changing datasets; and how we can use these models to help automate the production of software, which is capable of processing large volumes of data.

ALIGNED COMPETITION

As part of this workshop we will be running a special competition. We will offer 1 to 2 weeks of free consultancy services and training to the audience member who approaches us with the most interesting business problem. This is part of the ALIGNED business consultancy program.

Acknowledgements

The **ALIGNED project** (http://aligned-project.eu), the organisers of this workshop, is funded under grant agreement No 644055 of the European Union's Horizon 2020 research and innovation program. The organisation of this workshop has been supported by the **Cloudwatch project** (http://www.cloudwatchhub.eu) and the **Cluster on Software Engineering for Services and Applications** of the European Union's Horizon 2020 research program.

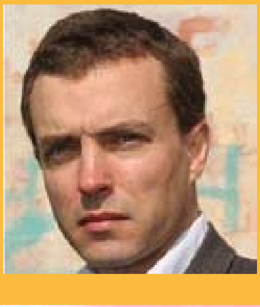
WORKSHOP SCHEDULE

10:00	Re-engineering a complex relational database application at Wolters Kluwer
10:45	Dacura – collecting and curating high quality datasets
11:30	Refreshments
11:30	Managing Data for the NHS
12:15	Integrating semantic datasets into Enterprise Information Systems with PooParty
13:00	Lunch
13:15	Moving Data Validation Closer to the Source at Wolters Kluwer

Wolters Kluwer – Re-engineering a complex relational database application

Christian Dirschl (Chief Content Architect), Wolters Kluwer Germany

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 actually 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, which is 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 quite some 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 also to be better prepared for any re-engineering effort in the future.



Dacura – collecting and curating high quality datasets

Dr. Kevin Feeney (Coordinator, ALIGNED) & Dr. Gavin Mendel Gleason (Integration Leader, ALIGNED) School of Statistics & Computer Science, Trinity College Dublin

In this presentation we will showcase the Dacura system – a platform which combines automated extraction of information from web sources with crowd-sourcing to radically cut the cost of collecting and curating high-quality datasets in any domain. We will use 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 will 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. This presentation should be of interest to anybody who wishes to do so quicker and cheaper than can be achieved using current technologies.

Managing Data for the NHS

Jim Davies (Professor of Software Engineering) University of Oxford

In this presentation, we will introduce 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 is to be processed, transformed, and accessed. Other tools support the processes of metadata capture and curation, as well as system configuration and generation.

We will explore the application of the Data Catalogue system to the management of health data in the UK. 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

Andreas Koller (CIO) & Robert David (CTO), Semantic Web Company Vienna

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, Semantic Web Company's semantic technology suite. PoolParty Thesaurus Server is a Thesaurus & Taxonomy Management Tool to build and maintain information architectures.

Most semantic web datasets have no fixed schema like relational databases, which may cause problems when importing these datasets into semantic web applications like PoolParty. PoolParty requires the imported datasets to conform its internal data structures. The PoolParty import assistant can be used to validate the correctness of the data and offers repair strategies for data curation. In this presentation we will 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.

This presentation should be of interest to anybody who wishes to increase the quality and consistency of semantic web data and linked data sets.

Moving Data Validation Closer to the Source at Wolters Kluwer

Markus Freudenberg (Release Manager), 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 presentation we will showcase how data validation can be far more effective when, instead of validating your instance data, you validate the application logic that generates and semantically transforms your data. That application logic can take the form of e.g. a custom application, an XSLT script or, a more advanced mapping language.

In a common scenario, we apply our approach to the Wolters Kluwer's Germany production data pipeline. We integrate it in their continuous integration (CI) system and measure improvement in productivity, agility and quality. We will additionally show, that in research and production settings with more sophisticated mapping languages, we can even infer logical inconsistencies directly from the mapping document. With this approach, we managed to reduce the validation time from 16 hours down to 30 seconds for very large datasets.