COMBUST

Creating a Platform to Fuse, Enrich, and Disseminate Reliable Business Data

The opportunities of big data are tremendous, be it for individuals as help to navigate today's digital environment, or for companies and authorities to underpin knowledgeable decisions. That makes offering data as a service an attractive business case. In addition, companies might add value to their data by enriching them through other data sets or crowdsourced data.

However, it is still a major challenge to combine available data from various heterogeneous sources and put it on offer in a way that is reliable, trustworthy and a good business case for the data owners.

The COMBUST project was set up to create solutions and guidelines for this data fusion challenge, in a realistic setting and in view of offering a valuable data service (DaaS – Data-as-a-Service).

"The challenges we run into when we fuse data from various parties and open it up are twofold," says Erik Mannens, research lead and professor at imec – IDLab – UGent. "First, there is the issue of data management. For every piece of data that is consulted, we need to know who is the owner and what is the level of trust we have in that data. In addition, the parties in such a data community have to make mutual agreements, e.g. agreeing on the fees they ask to make their data available. Second, there is the technical issue of merging data, versioning, enriching... in a scalable way, so that e.g. additional data partners can join without causing a major overhead for everyone involved."

The COMBUST partners allow implementing a realistic and complete business scenario. They include three content providers that own valuable but non-competitive data sets that they each can enrich with the data of the other two. In addition, there is a communication provider who can draw on a large user base to organize data crowdsourcing to further enrich the data sets.

And last, there are a solution provider and big data researchers with the expertise to implement scalable, high-quality tools and guidelines.

THE OUTCOMES

1. Creating a scalable flow to merge and enrich data

As a basis to merge data from various sources and formats, we've chosen an open intermediate format (RDF) that has proven its worth in many large-scale projects around the world. We have set up a user-friendly mapping editor that allows each party to define how their own data will be mapped onto the common RDF format. This is key to make the model scalable: for each additional party joining the effort, only one new mapping has to be made.
The mapping processor then takes care of the actual mapping, adding information about the data's provenance. Once the data is mapped, it is published in a format called Linked Data Fragments. This is a format and technology that distributes the load of querying data sets between the server and client, adding another element of scalability. These data fragments can now merged and enriched, after which they may be remapped to the parties' own data sets.

2. Adding the ability for intelligent crowdsourcing

People who are using data, e.g. though their smartphones, are often willing to correct and enrich data. They may e.g. add info about actual opening hours, waiting times, location... That is why the project also developed a dedicated app, nicknamed Odin, to query user panels. It directly interfaces to the Linked Data Fragments. And using the available semantic information, it is able to automatically formulate questions for the app users.

3. Adding trust and business value with versioning

Data providers want to know where the data they add to their sets originates and if it is trustworthy. Also, as part of a business agreement and payment scheme, they want to keep track of who owns the data that they make available. That is why COMBUST has added a powerful versioning tool to its flow, the first such tool created and used outside of an academic context.

NEXT STEPS

When fully developed, COMBUST has a good chance of becoming a unique collaborative environment which creates external benefits, internal revenue and the perspective to be a player on the international DaaS market that is flexible and growth-oriented enough to play an important role globally.

COMBUST project partners:

The COMBUST project was co-funded by imec (iMinds), with project support from Agentschap Innoveren & Ondernemen.

FACTS

NAME	COMBUST (Combining Multi-trUst BusinesS daTa)
OBJECTIVE	The project will implement the tools to create an enriched knowledge framework based on existing and new data sources. Additionally, it will examine the potential of Data-as-a-Service (DaaS), exploring the various business models and opportunities.
TECHNOLOGIES USED	RDF, RML, Linked Data, Linked Data Fragments, Provenance
TYPE	imec.icon project
DURATION	01/01/2015 - 31/03/2017
PROJECT LEAD	Roel De Meester, Thanksys
RESEARCH LEAD	Erik Mannens, imec – IDLab – UGent
BUDGET	2,354,970 euro
PROJECT PARTNERS	GIM NV, Truvo (nu FCR Media Belgium), Roularta Business Information, Thanksys, Mobile Vikings
IMEC RESEARCH GROUPS	IDLab – UGent, imec.livinglabs, MICT – UGent, SMIT – VUB

WHAT IS AN IMEC.ICON PROJECT?

The imec.icon research program equals demand-driven, cooperative research. The driving force behind imec.icon projects are multidisciplinary teams of imec researchers, industry partners and / or social-profit organizations. Together, they lay the foundation of digital solutions which find their way into the product portfolios of the participating partners.