The BigIaS Platform

Simplifying Big Data Integration

- A Software-as-a-Service Approach –

~ Preliminary Analysis and Design ~

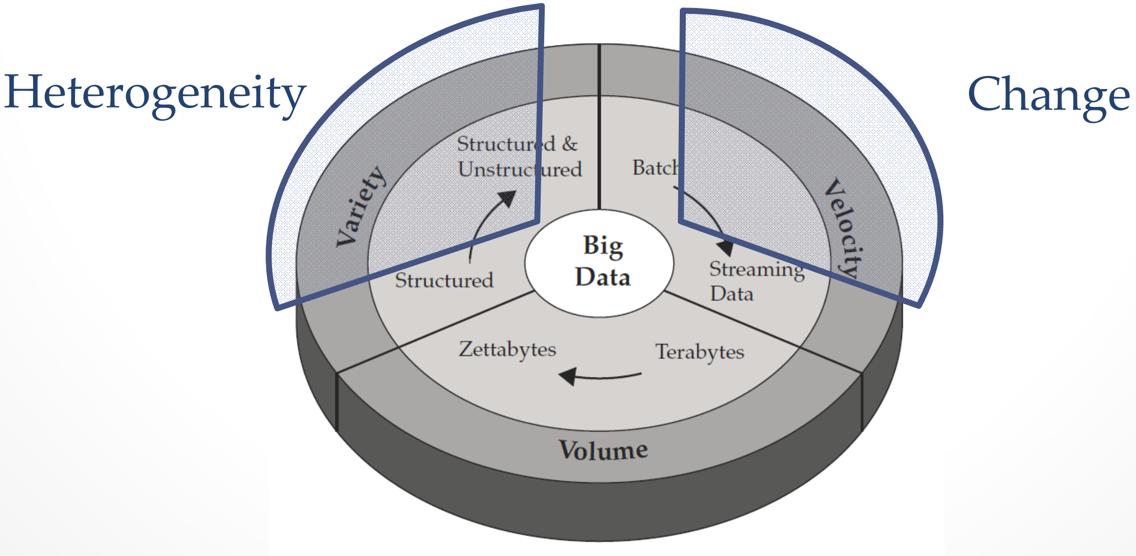
September 5th, 2013
Sogndal, Norway

Dumitru Roman
Claudia Daniela Pop
Roxana Ioana Roman
Bjørn Magnus Mathisen

Contact: dumitru.roman@sintef.no

Context: Big Data and Our Primary Focus

- Addresses things that can be done at a large scale but cannot be done at a smaller one
  - Extract new insights or create new forms of value in ways that change stakeholders and relationships between them
- Causality vs. correlations: not knowing *why* but only *what*
  - Challenges the basic traditional understanding of how to make decisions



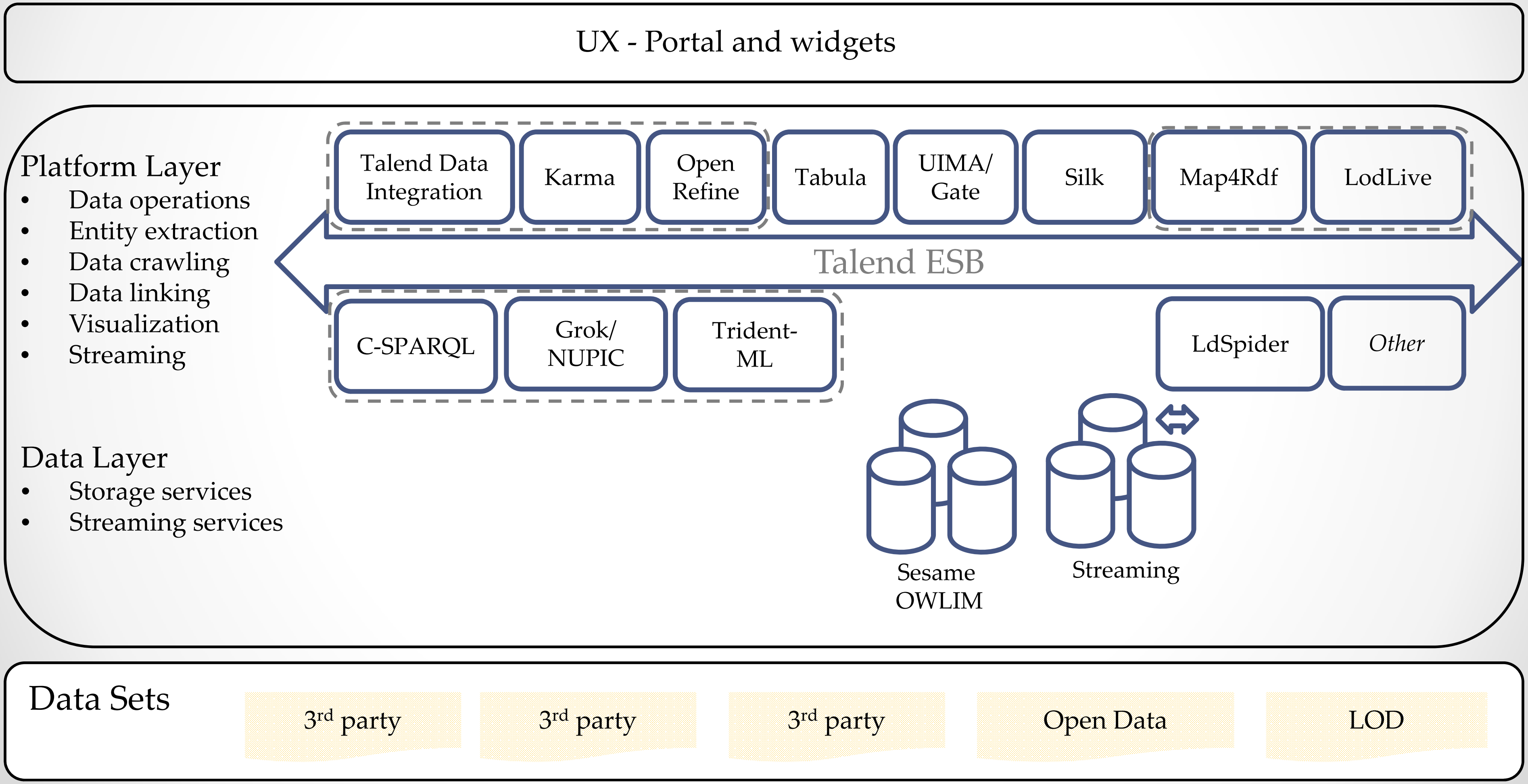
Overview

- The problem
  - Data integration - a complex, unsolved problem
  - Tools addressing various aspects of data integration process can hardly be used together for more complex, interesting integration tasks

  => High cost of data integration at large scale, rather complicated and time consuming process
- The goal: Simplify data integration at large scale!
  - Enable users with limited technical data integration skills to get from raw data to insightful data with minimal effort
- The approach
  - Semantic-based data integration
  - Flexible and customizable workflows of data integration tools (application integration)

  => A Software-as-a-Service for data integration at large scale

The BigIaS Platform



UX - Portal and widgets

Platform Layer
- Data operations
- Entity extraction
- Data crawling
- Data linking
- Visualization
- Streaming

Data Layer
- Storage services
- Streaming services

Data Sets

Evaluated tools/approaches

1. **Application Integration**
   - Talend ESB
2. **Data Processing**
   - Talend Data Integration
   - Tabula
   - Karma
   - Open Refine
   - UIMA
   - GATE
   - Silk
   - LdSpider

3. **Storage**
   - Sesame
   - OWLIM
4. **Visualization**
   - Map4Rdf
   - LodLive

5. **Real-time Machine Learning**

- Trident-ML
- Grok/NUPIC

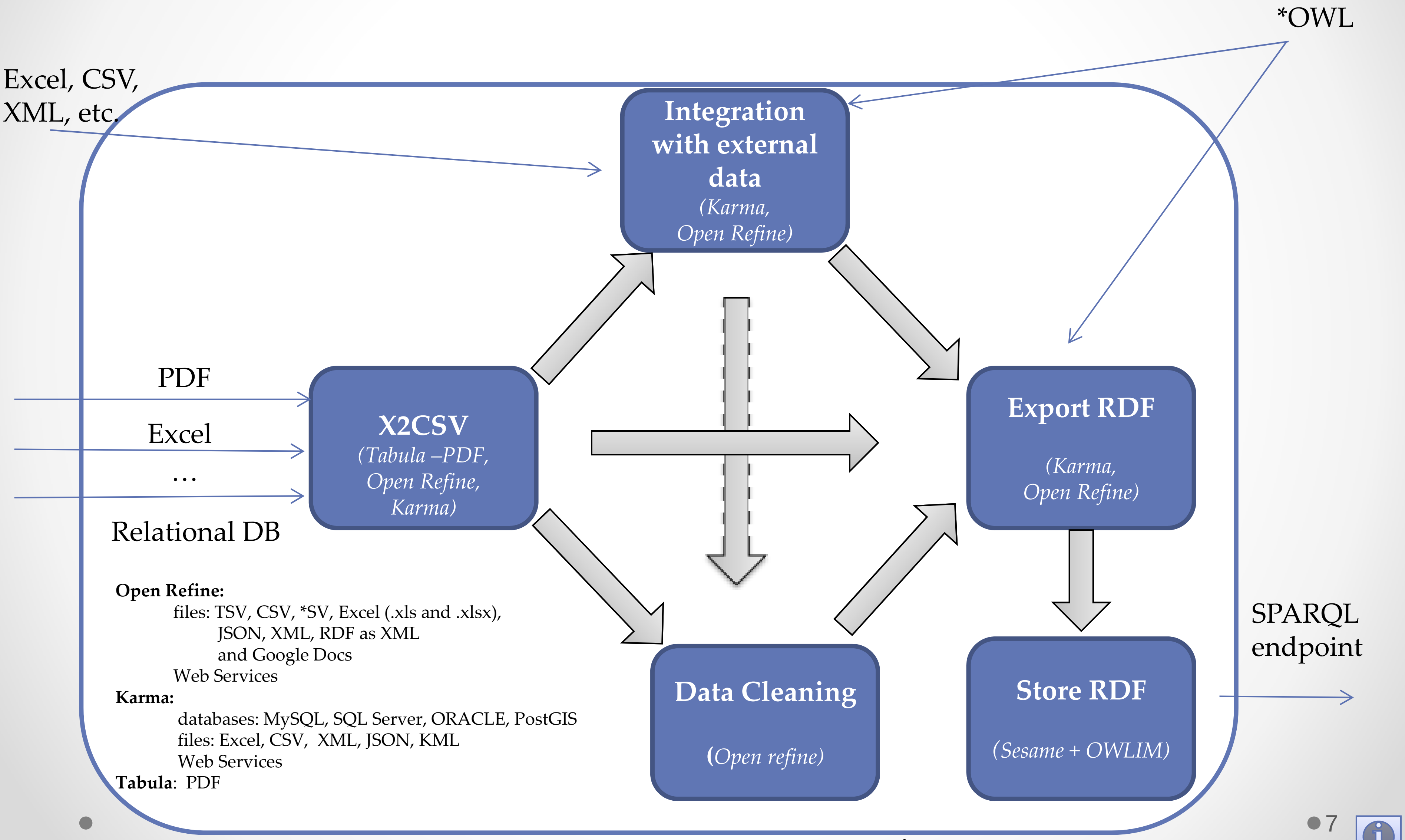
5. **Streaming**

- C-SPARQL

Proposed Integration Workflows

1. Data Contextualization
2. Entity Discovery
3. Data Linking
4. Data Visualization
5. Real-time Machine Learning
6. RDF Streaming

Data Contextualization



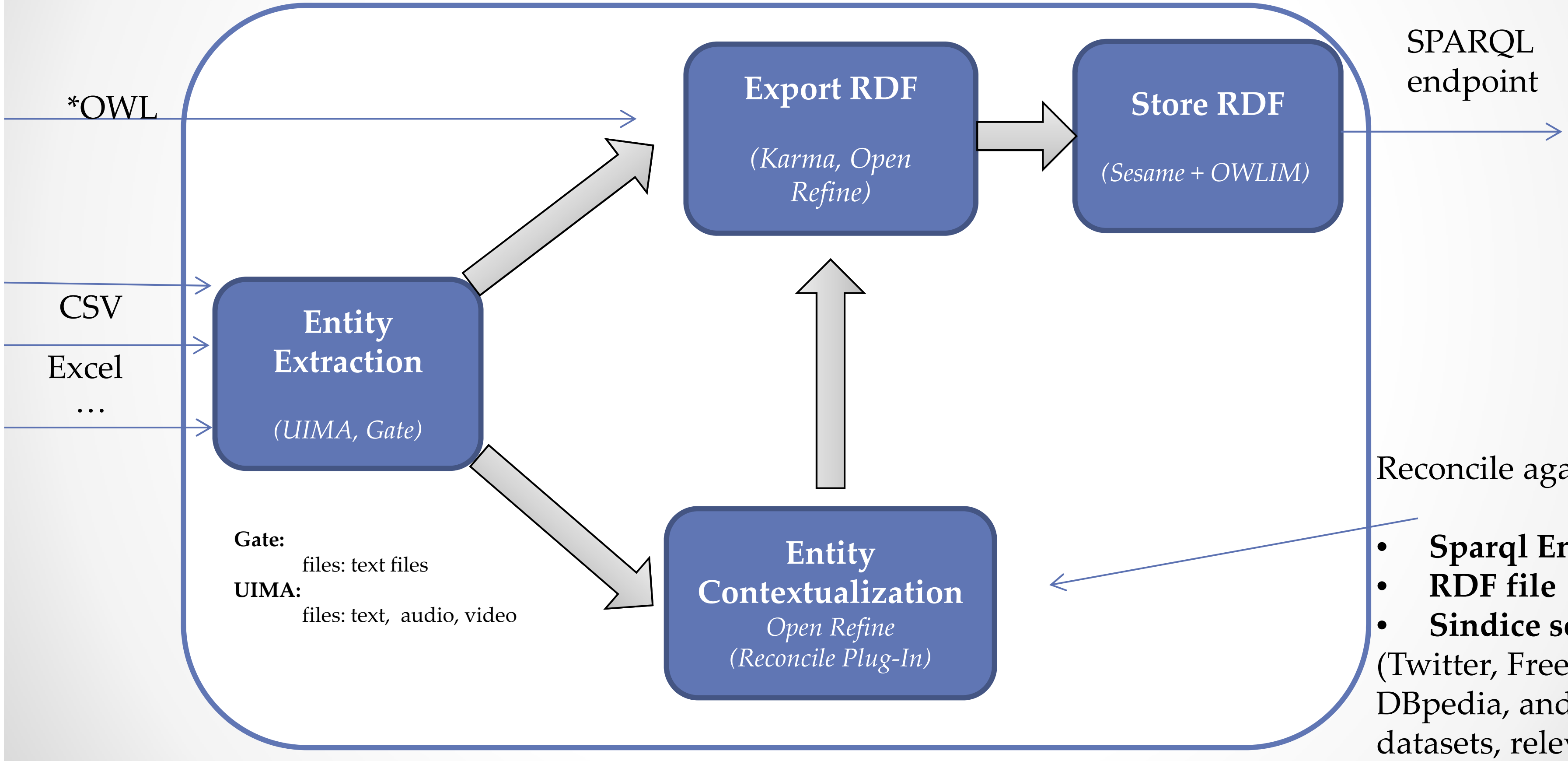
Open Refine:
files: TSV, CSV, *SV, Excel (.xls and .xlsx),
JSON, XML, RDF as XML
and Google Docs
Web Services

Karma:
databases: MySQL, SQL Server, ORACLE, PostGIS
files: Excel, CSV, XML, JSON, KML
Web Services

Tabula: PDF

* OWL file can be imported or configured

Entity Discovery



*OWL

CSV

Excel

…

SPARQL endpoint

Export RDF

(Karma, Open Refine)

Store RDF

(Sesame + OWLIM)

Entity Extraction

(UIMA, Gate)

Entity Contextualization

Open Refine (Reconcile Plug-In)

Gate:
files: text files

UIMA:
files: text, audio, video

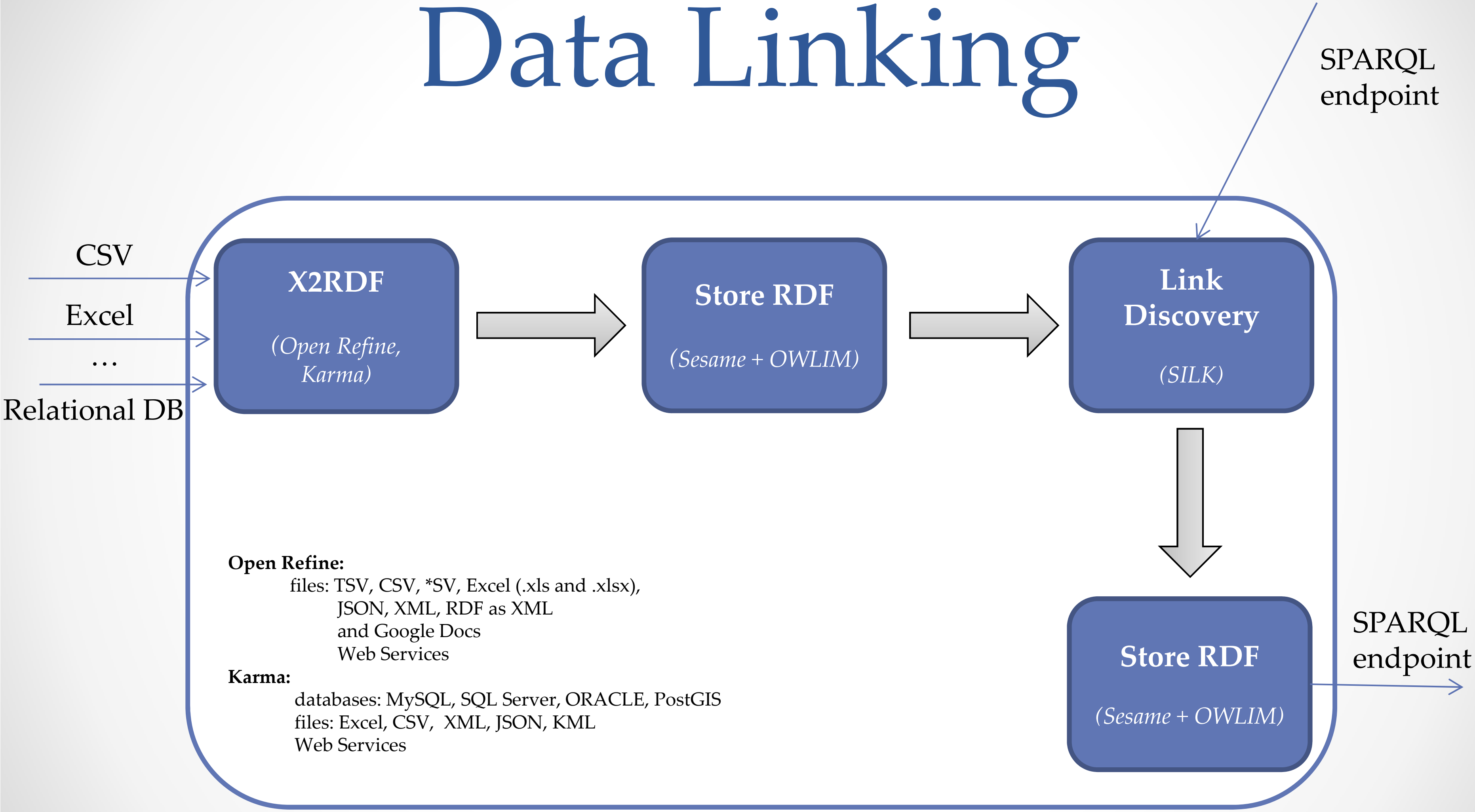
Reconcile against:

- **Sparql Endpoint**
- **RDF file**
- **Sindice service**

(Twitter, Freebase, DBpedia, and other datasets, relevant for the data)

* OWL file can be imported or configured

Data Linking



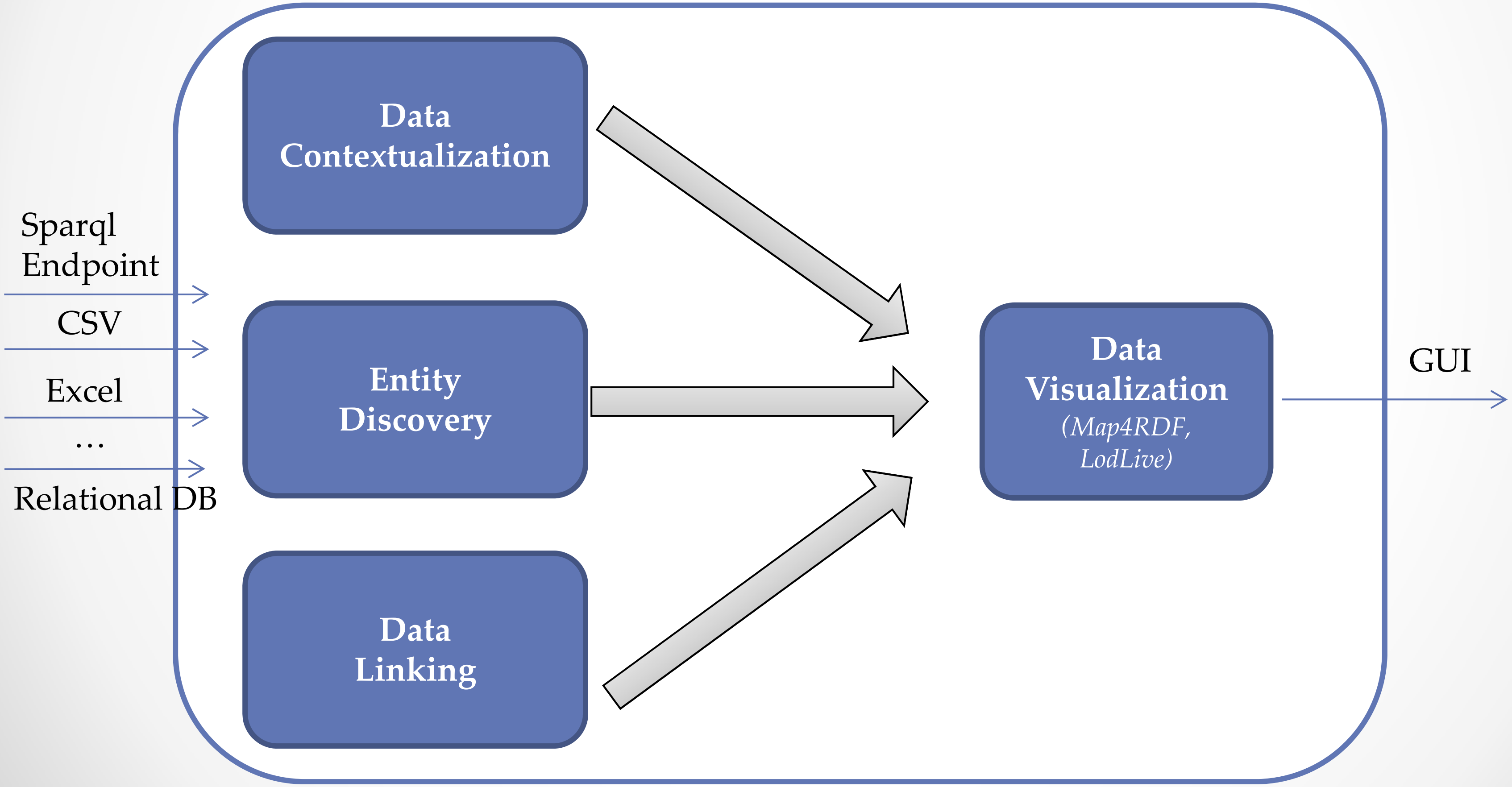
Open Refine:

files: TSV, CSV, *SV, Excel (.xls and .xlsx),
JSON, XML, RDF as XML
and Google Docs
Web Services

Karma:

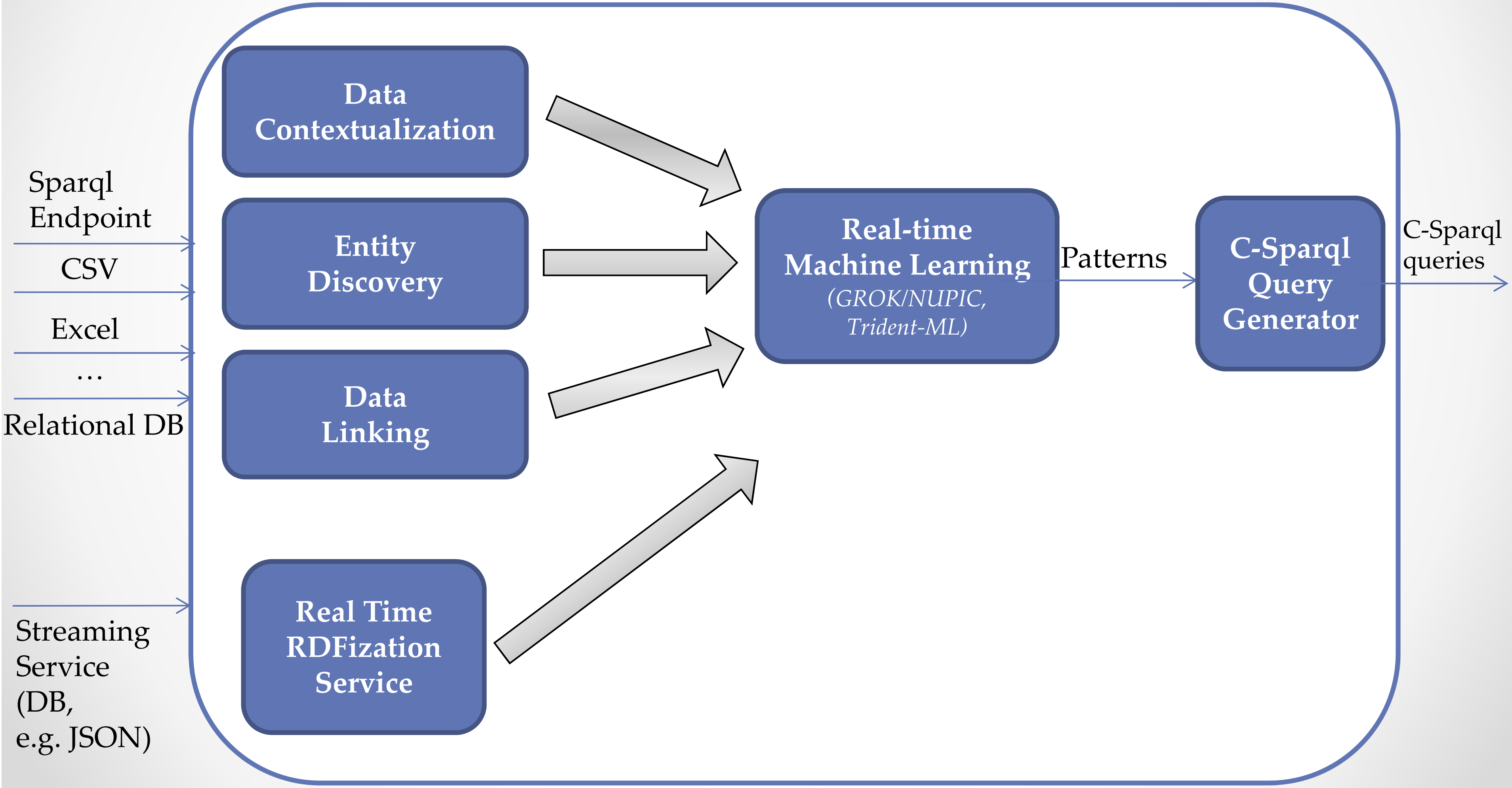
databases: MySQL, SQL Server, ORACLE, PostGIS
files: Excel, CSV, XML, JSON, KML
Web Services

Data Visualization

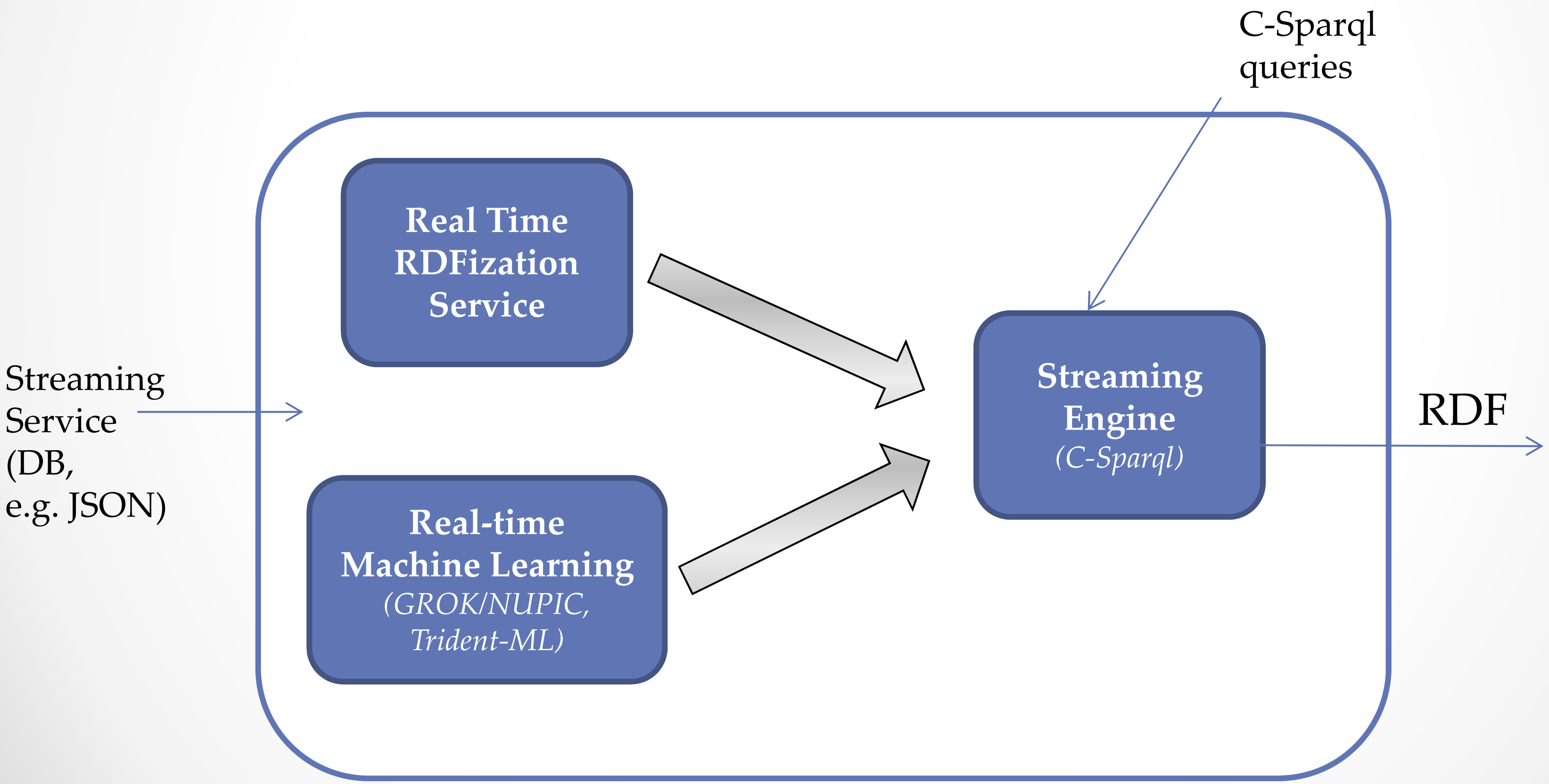


SINTEF

Real-time Machine Learning



RDF Streaming



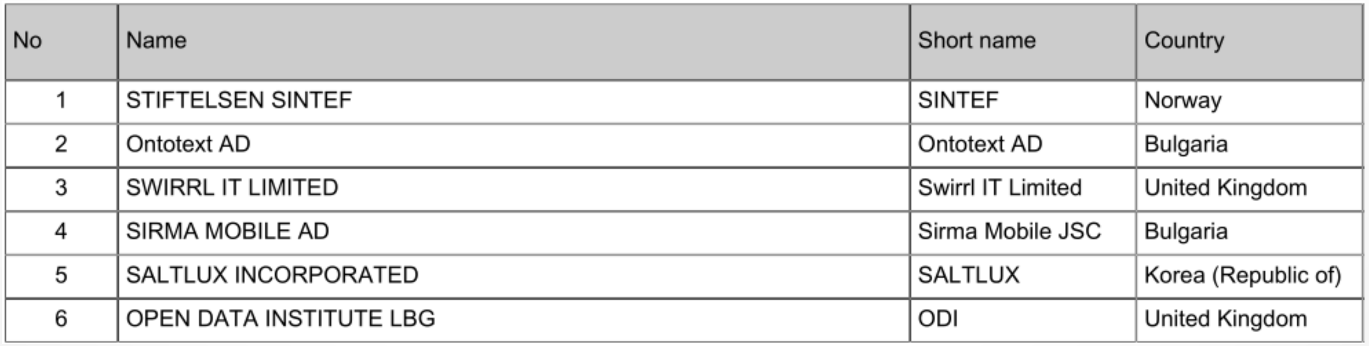
Core Technological Prerequisites

Tool	Prerequisites
Open Refine	GREL Functions GREL Examples
Karma	Create ontologies Ontologies from Karma Web Interface
UIMA	Regular Expressions
Silk	Silk Link Specification Language (Silk-LSL)
C-Sparql	C-Sparql language

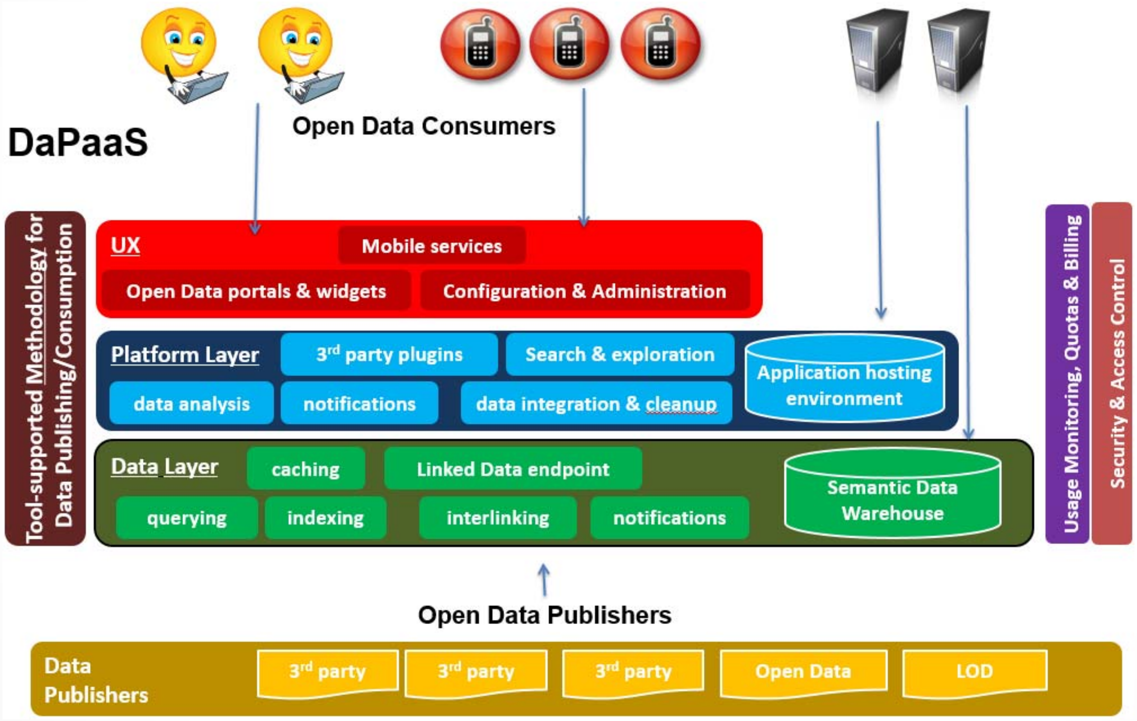
Relevant upcoming research projects

(currently under negotiations)

- *Data Publishing through the Cloud: A Data- and Platform-as-a-Service Approach for Efficient Data Publication and Consumption (DaPaaS)*
  - *The DaPaaS project aims to deliver an integrated DaaS and PaaS environment for open data–the DaPaaS platform–together with supporting activities for effective and efficient publication and consumption of data and creation of applications using the data.*
  - Expected to start Nov 2013
  - Budget ~2.1M € (~1.5M €) for 2 years (EC funded)



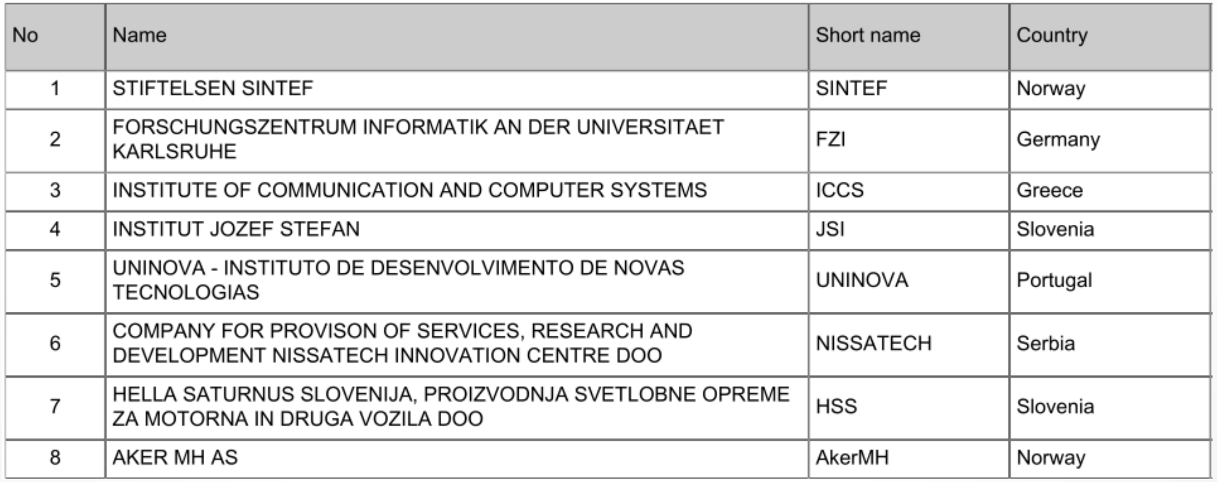
No	Name	Short name	Country
1	STIFTELSEN SINTEF	SINTEF	Norway
2	Ontotext AD	Ontotext AD	Bulgaria
3	SWIRRL IT LIMITED	Swirrl IT Limited	United Kingdom
4	SIRMA MOBILE AD	Sirma Mobile JSC	Bulgaria
5	SALTLUX INCORPORATED	SALTLUX	Korea (Republic of)
6	OPEN DATA INSTITUTE LBG	ODI	United Kingdom


DaPaaS
Open Data Consumers
Tool-supported Methodology for Data Publishing/Consumption
UX
Mobile services
Open Data portals & widgets
Configuration & Administration
Platform Layer
3rd party plugins
Search & exploration
Application hosting environment
data analysis
notifications
data integration & cleanup
Data Layer
caching
Linked Data endpoint
Semantic Data Warehouse
querying
indexing
interlinking
notifications
Usage Monitoring, Quotas & Billing
Security & Access Control
Open Data Publishers
Data Publishers
3rd party
3rd party
3rd party
Open Data
LOD

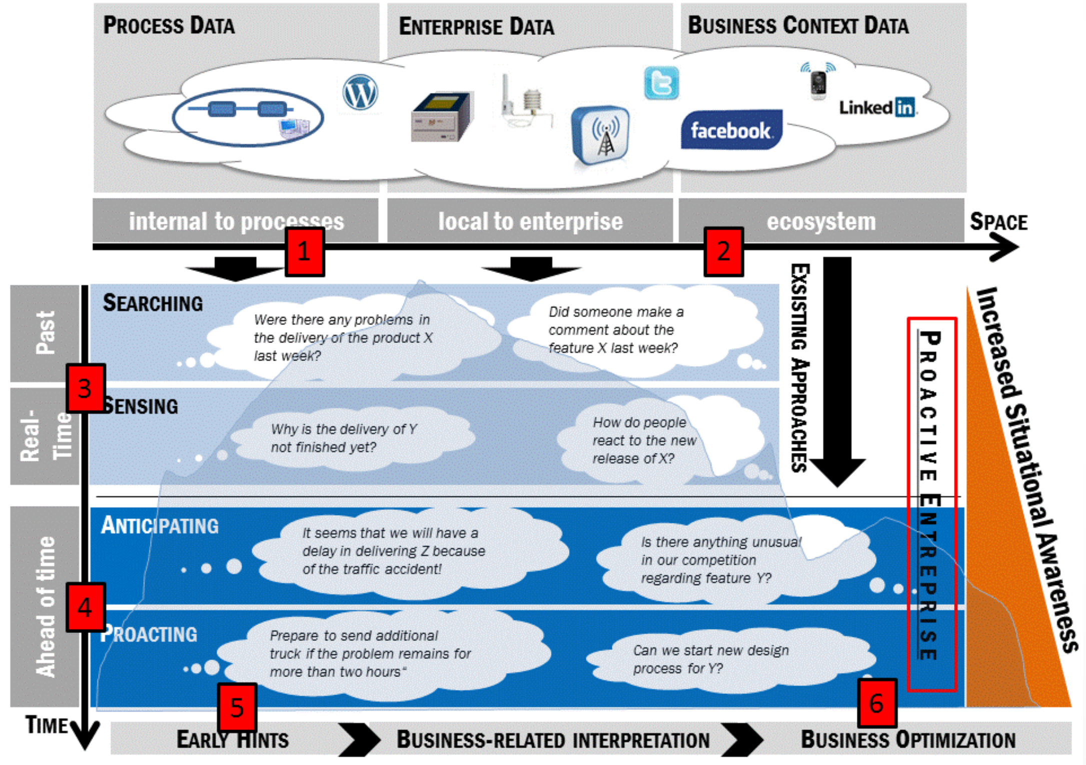
Relevant upcoming research projects

(currently under negotiations – cont')

- ProaSense – The Proactive Sensing Enterprise
  - The goal is to provide a very scalable, distributed architecture for the management and processing of big-data that will enable continuous monitoring of the need for the service adaptation and propose corresponding changes in an (semi-) automatic way.
  - Expected to start Nov 2013
  - Budget ~4.2M € (~3.2M €) for 3 years (EC funded)

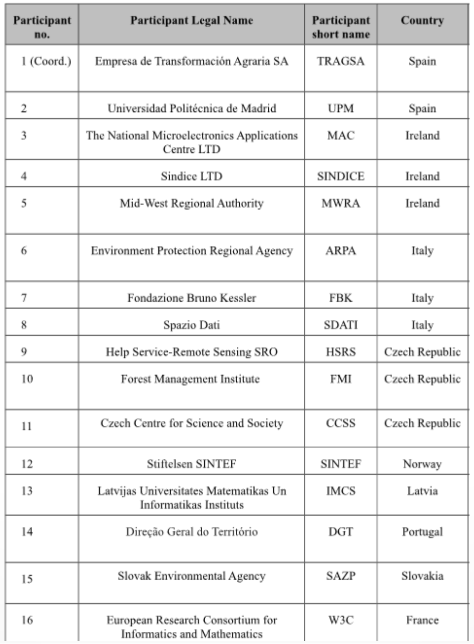


No	Name	Short name	Country
1	STIFTELSEN SINTEF	SINTEF	Norway
2	FORSCHUNGSZENTRUM INFORMATIK AN DER UNIVERSITAET KARLSRUHE	FZI	Germany
3	INSTITUTE OF COMMUNICATION AND COMPUTER SYSTEMS	ICCS	Greece
4	INSTITUT JOZEF STEFAN	JSI	Slovenia
5	UNINOVA - INSTITUTO DE DESENVOLVIMENTO DE NOVAS TECNOLOGIAS	UNINOVA	Portugal
6	COMPANY FOR PROVISON OF SERVICES, RESEARCH AND DEVELOPMENT NISSATECH INNOVATION CENTRE DOO	NISSATECH	Serbia
7	HELLA SATURNUS SLOVENIJA, PROIZVODNJA SVETLOBNE OPREME ZA MOTORNA IN DRUGA VOZILA DOO	HSS	Slovenia
8	AKER MH AS	AkerMH	Norway


PROCESS DATA
ENTERPRISE DATA
BUSINESS CONTEXT DATA
facebook
Linked in
internal to processes
local to enterprise
ecosystem
SPACE
1
2
EXSISTING APPROACHES
PROACTIVE ENTERPRISE
Increased Situational Awareness
Past
Real-Time
Ahead of time
3
4
SEARCHING
Were there any problems in the delivery of the product X last week?
Did someone make a comment about the feature X last week?
SENSING
Why is the delivery of Y not finished yet?
How do people react to the new release of X?
ANTICIPATING
It seems that we will have a delay in delivering Z because of the traffic accident!
Is there anything unusual in our competition regarding feature Y?
PROACTING
Prepare to send additional truck if the problem remains for more than two hours"
Can we start new design process for Y?
5
6
TIME
EARLY HINTS
BUSINESS-RELATED INTERPRETATION
BUSINESS OPTIMIZATION

Relevant upcoming research projects
(currently under negotiations – cont')

- SmartOpenData – Open Linked Data for environment protection in Smart Regions
  - SmartOpenData aims to define mechanisms for acquiring, adapting and using Open Data provided by existing sources for environment protection in European protected areas
  - Expected to start Nov 2013
  - Budget ~3.4M € (~2.5M €) for 2 years (EC funded)

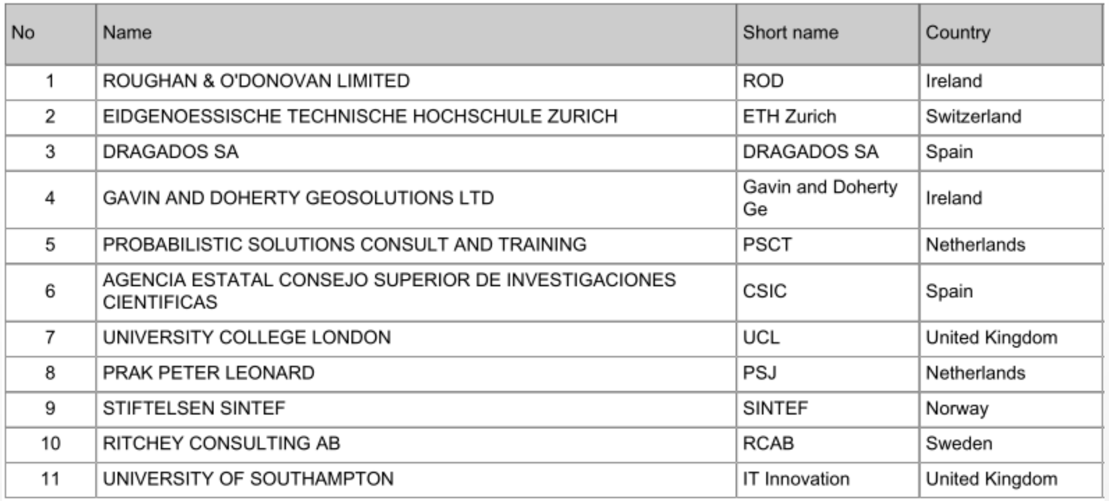


Participant no.	Participant Legal Name	Participant short name	Country
1 (Coord.)	Empresa de Transformación Agraria SA	TRAGSA	Spain
2	Universidad Politécnica de Madrid	UPM	Spain
3	The National Microelectronics Applications Centre LTD	MAC	Ireland
4	Sindice LTD	SINDICE	Ireland
5	Mid-West Regional Authority	MWRA	Ireland
6	Environment Protection Regional Agency	ARPA	Italy
7	Fondazione Bruno Kessler	FBK	Italy
8	Spazio Dati	SDATI	Italy
9	Help Service-Remote Sensing SRO	HSRS	Czech Republic
10	Forest Management Institute	FMI	Czech Republic
11	Czech Centre for Science and Society	CCSS	Czech Republic
12	Stiftelsen SINTEF	SINTEF	Norway
13	Latvijas Universitates Matematikas Un Informatikas Instituts	IMCS	Latvia
14	Direção Geral do Território	DGT	Portugal
15	Slovak Environmental Agency	SAZP	Slovakia
16	European Research Consortium for Informatics and Mathematics	W3C	France

Relevant upcoming research projects

(currently under negotiations – cont')

- INFRARISK— Novel Indicators for identifying critical INFRAstructure at RISK from natural Hazards
  - Develop reliable stress tests on European critical infrastructure using integrated modelling tools for decision-support. It will lead to higher infrastructure networks resilience to rare and low probability extreme events, known as "black swans".
  - Expected to start Oct 2013
  - Budget ~3.6M € (~2.8M €) for 2 years (EC funded)



No	Name	Short name	Country
1	ROUGHAN & O'DONOVAN LIMITED	ROD	Ireland
2	EIDGENOESSISCHE TECHNISCHE HOCHSCHULE ZURICH	ETH Zurich	Switzerland
3	DRAGADOS SA	DRAGADOS SA	Spain
4	GAVIN AND DOHERTY GEOSOLUTIONS LTD	Gavin and Doherty Ge	Ireland
5	PROBABILISTIC SOLUTIONS CONSULT AND TRAINING	PSCT	Netherlands
6	AGENCIA ESTATAL CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS	CSIC	Spain
7	UNIVERSITY COLLEGE LONDON	UCL	United Kingdom
8	PRAK PETER LEONARD	PSJ	Netherlands
9	STIFTELSEN SINTEF	SINTEF	Norway
10	RITCHEY CONSULTING AB	RCAB	Sweden
11	UNIVERSITY OF SOUTHAMPTON	IT Innovation	United Kingdom

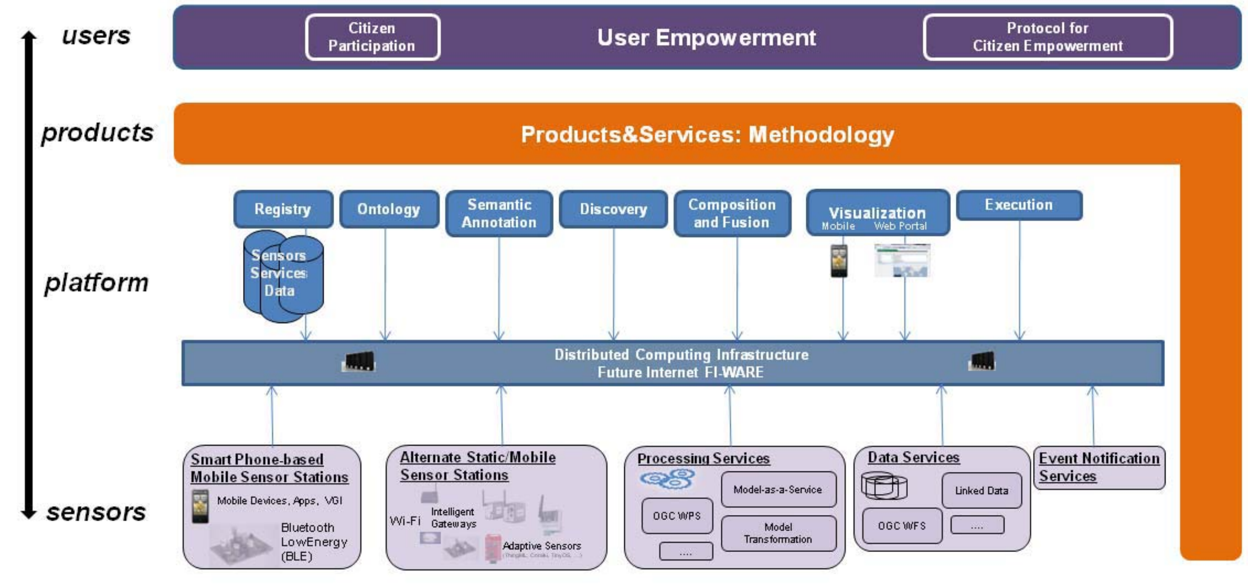
Relevant ongoing research projects

- BigFut – Analyzing Big Data: Preparing for the Future of Intelligent Information Management
  - SINTEF Internal project
  - Goals:
    - Analyze and integrate a suit of technological approaches and techniques
    - Advance demo/prototype implementation to penetrate the market in the short term
  - Jan-Dec 2013

Relevant ongoing research projects (cont')

- CITI-SENSE – develop "Citizen's Observatories" to empower citizens to:
  - Contribute to and participate in environmental governance
  - Support and influence community and policy priorities and associated decision making
  - Contribute to Global Earth Observation System of Systems (GEOSS)
- 27 partners (EC fuded)

http://www.citi-sense.eu/



Summary and Outlook

- What's new here
  - The platform itself, implementing flexible data integration workflows
  - A set of components (e.g. Real-time RDFization of streams, C-Sparql Query Generator)

- What's challenging
  - Application integration
  - Consistent scalability throughout workflows
  - Platform deployment on cloud environments
  - Use of new, unproven technologies
  - …and many others

- Short-term plan (end of September)
  - Get the first prototype implementing the proposed workflows
  - Experiment with some data / simple use cases (e.g. CITI-SENSE data)

Thank you!

Q&A