



Co-funded by the European Commission



# Annual Public Report 2009

The main objective of the OKKAM R&D project is to enable the **Web of Entities**, which is an “entity-centric” layer on top of the current web where new applications and services can be enabled based on the fact that entities are uniquely identified across any type of format, language, application.

To this end, the OKKAM Consortium will deploy a public global service, called **Entity Name System** (ENS for short), which any application can call to retrieve a uniquely global ID for any type of entity, e.g. people, organizations, locations, events, named in an electronic information source or document and create customized annotations. Several applications for managing data and content will be enabled (typically through plugins) to interact with the ENS and produce entity-centric data and content.

The potential of this idea will be demonstrated in three critical areas: *enterprise knowledge management, publishing, and web search*. Other smaller pilot projects have been started in areas like *eGovernment* and *Public Administration*.

The OKKAM Project involves 12 partner organisations from 7 European countries and will end by June 2010.

## 1 Summary of Activities

Along the second year of the project, the OKKAM Consortium has concentrated its efforts in four main directions:

- Implementing the second prototype of the ENS and making it available to third parties for dissemination and demonstration purposes. This enhanced ENS is designed to be highly scalable, to have better performance and to offer a first level of access control and protection of data.
- Extending the collection of OKKAM-empowered tools, which now include plugins for MS Word, OpenOffice, Outlook, Protégé, NeOn Toolkit, Gmail, Hotmail, Blogger, Mozilla Firefox and Internet Explorer. In addition to these tools, we offer an OKKAM-enabled web interface for creating Friend-Of-A-Friend (FOAF) profiles and a web-based tool for research and experimenting the process of annotating text with OKKAM identifiers for entities
- Designing the application prototypes empowered by OKKAM technologies for fulfilling the identified business needs of the OKKAM application partners (SAP, Elsevier, ANSA).
- Creating the bases for future research and exploitation of OKKAM outputs. On the one hand, by progressively creating a “critical mass” of scientific and industrial organizations potentially interested in using the ENS and related technologies. And on the other, designing and starting to implement exploitation strategies. On top, the OKKAM consortium created a Community Portal (<http://community.okkam.org/>) which provides a unified access point for researchers, potential developers and end users.

Achievements witnessed so far allows the understanding that OKKAM can offer valuable solutions to those challenges by improving their performance through a better exploitation of the opportunities created by the pervasiveness of the Internet.

## 2 Important Work Areas

In the second year, the OKKAM consortium focused on the following work areas.

### 2.1 Storage layer & indexing

The ENS requires a very large and scalable infrastructure to serve a large number of requests per second and to store a huge amount of data (potentially about billions of entities). This means that we need to ensure very high scalability. In the second year, the effort was to further improve the storage and indexing infrastructure. In the latest version of the ENS (2.0), both the index and the storage is distributed, in order to achieve better performance figures, improve scalability and ensure higher availability through replication. The storage and indexing component relies on open source projects, such as Solr, lucene, Hbase, Project voldemort,

but further extends and adapts them along the requirements of the ENS. These adaptations include specific entity-oriented ranking, computing of entity-specific statistics. The storage layer serves as an underlying infrastructure and provides specific services for the other components within the ENS.

## **2.2 *Advanced Entity Matching methods***

One of the key challenges in OKKAM is to design and develop methods and techniques which can be used to decide whether a piece of data describing an entity corresponds to an entity already stored in the OKKAM entity repository without presupposing a fixed schema for representing information about stored entities and for a wide variety of entity ID request types.

In the second year, the OKKAM team focussed on three core areas: first, the matching functionality was enhanced with full probabilistic semantics and domain specific similarity functions for the request language, which is used to search for entity identifiers. Second, a test suite was designed which systematically and flexibly supports automated large-scale experiments. Third, we have investigated advanced entity search related issues, which include representing and modeling the uncertainties that may appear in the matching process. Furthermore, to improve matching effectiveness, the OKKAM team started investigating the techniques for extending entity matching methods with knowledge from external sources (e.g., WordNet) as well as knowledge resulting from examining inner-relationships across entities.

The most recent activities include the extension of functionality with bulk entity import, which will allow applications to more efficiently connect and use the OKKAM matching process and search functionality on their own data.

## **2.3 *Lifecycle Management***

The focus of the Entity Lifecycle Management (ELM) component is on the management of the information in the ENS repository for a particular entity, throughout the lifetime of that entity in the repository. As such, ELM lies at the core of the ENS. More specifically, the ELM component deals with issues that arise in entity representation, data quality, repository evolution, and online monitoring of the use of the repository.

The ELM component provides functionality that cover the basic operations of creating, updating or modifying, and deleting (only for system administrators) entities. In representing an entity, our goal in ENS is to keep this representation as simple and general as possible, without imposing a fixed schema of entity types or attributes. We take special care in assuring data quality at creation time, as well as over the lifetime of the entity representation in the repository. We support the operations of entity merge and split, which may be necessary we identify that two different entity representations refer to the same real world entity, or when the

same entity representation is already being used to refer to two distinct real world entities. We monitor and analyze the way that users access the system and interact with it, and use this information in order to improve the performance of the ENS. Finally, through the ELM component we offer mechanisms for personalized user-notifications on changes in the entity descriptions, which facilitate user involvement in the process of maintaining the information in the ENS repository.

## **2.4 Security**

Security in OKKAM is focused on controlling access to the information stored on the ENS with the objectives of protecting privacy-sensitive information on both ENS and users' sides. Security mechanisms are designed to be flexible to cope with information heterogeneity, transparent to foster wide usage, and efficient to support ENS expectations of high throughput.

More technically, OKKAM security architecture is based on advanced use of certificate technologies, beyond the classic notion, with support of decentralized credential management and enhanced user control over authorization processes.

An additional feature of the security architecture is to support the strategic goal of OKKAM for fostering the creation of community of users, by providing them with the full benefits of certificate technologies, such as capability of establishing confidential communications between end-users, and document authentication based on digital signatures.

The security architecture has been successfully implemented and deployed in the ENS 2.0 release by the following main components - certificate authorities for issuance and management of certificates, security proxies of ENS and users' side, and access control decision and enforcement components.

## **2.5 ENS 2.0: integration and deployment**

The second prototype of the ENS was completed and validated on the OKKAM Test Bed, and launched on the OKKAM Operational Platform in November, to allow public operation of the service. This major integration and upgrade of the ENS service to a distributed, secure and scalable architecture was achieved through distributed collaborative development by a Technical Task Force involving staff from many OKKAM Partners. The work required adherence to agreed development procedures, regular automated builds and validations of the system, very many virtual and actual meetings, and use of the OKKAM gForge development management environment. A Roadmap of ENS releases up to V3, for deployment at the end of the project, was developed. The ENS V2.0 prototype is now being used to dimension and characterise the service requirements for the permanent OKKAM Production Platform to be put in place during the last phase of the project.

### 3 User Involvement, Promotion and Awareness

User involvement is a key feature in OKKAM. Users are represented by well-defined Application Scenarios, instantiated directly by project partners:

**Enterprise Knowledge Management:** together with SAP, we are creating an entity-centric application which may support SAP customers to find answers to their problems and experts in the SAP community portals in a much faster and more precise way. The application will identify relevant entities (and their relations) in a user's request and try to match it against past answers to related requests about the same entity (e.g. a specific product) and against experts' profiles. The answer may include material produced outside the SAP community network (e.g. in an informal forum for developers), as external material can also be annotated with OKKAM identifiers.

**Authoring and Publishing:** together with our partners Elsevier and ANSA, we are creating an authoring environment in which authors (specifically, scientists and journalists) can be supported in creating their content in a faster and better way, and to collect additional information on what they write in a semi-automated way by exploiting the recognition of known entities in their text. Examples can be: proteins in biological papers, past events in news items.

**Search:** together with DERI Ireland, we are building an entity-centric semantic search engine, called Sigma, on top of the Sindice search engine (<http://www.sindice.com>). The current version is available at <http://sig.ma/>. Sigma can be used to send a query about a given entity (e.g. the European Commission) and the result is not a list of documents, but a structured profile of the requested entity from which we can start for browsing the space of entities on the web. Sigma provides very useful functionalities, including: creating a permanent link to an entity profile, filtering out irrelevant results, viewing only statements from selected sources, generation of Javascript code for including entity profiles in regular web pages.

In addition to these use cases, the OKKAM consortium has started two pilot projects with external partners:

1. **eGovernment:** the project aims at using OKKAM identifiers and the ENS to tag physical objects in a city and redirect different applications to different information sources about the same object. The partner is the City of Manor (TX, USA).
2. **Data aggregation & mashup:** the project aims at automatically creating browsable profiles of tax payers in Trentino starting from a collection of many distributed heterogeneous sources. The partner is Trentino Riscossioni (a regional tax agency) and the government of the Province of Trento (Italy).

With respect to our activities aimed at promoting take-up of project outputs, several OKKAM-related scientific events were co-organized by members of the consortium: Semantic Web Applications and Perspectives Workshop (SWAP2009), the 2<sup>nd</sup> International Conference on Digital Libraries and Semantic Web (ICSD2009), the IJCAI-09 Workshop on Identity and Reference in Web-based Knowledge Representation (IR-KR2009).

Two special events were aimed at the investors' world. On the one hand, the Web 3.0 Venture Academy, mainly set up as an opportunity for the OKKAM consortium to meet with investors and gather feedback and input on OKKAM technologies, its strengths and weaknesses and on potentially important issues to be considered for the commercialisation strategy and business plan. On the other, the OKKAM concept and business plan was presented at the 1st Trentino Technology Tour. Closing the year, an OKKAM tutorial at the 3rd European Semantic Technology Conference (ESTC2009) will be organized in Vienna, Austria.

#### **4 Future Work and Exploitation Prospects**

The OKKAM project will enter 2010 facing the challenge to manage the completion of the Project and, at the same time, to create the conditions for the actual exploitation of the ENS infrastructure and other exploitable results.

Among particular endeavours to be completed in the next few months:

- the deployment of the **ENS on a production platform** and the completion of a plan for the next releases
- the consolidation of the **research results** and their dissemination in the appropriate channels and using exciting **showcases and eventually** small scale **demonstration projects**
- Implementing the masterplan towards the creation of the **OKKAM Foundation** which will collect resources for the operations and evolution of the ENS, and will guarantee the neutrality and independence of its usage.

The potential for synergies is huge. Therefore, we are looking for partners and initiatives which may benefit from adopting the ENS infrastructure as a global naming (or ID) system.

In this respect, a non-exhaustive list of interesting scenarios includes:

- Projects in which large collection of data are produced and exposed for machine consumption ("web of data", linked data)
- Web service-oriented projects in which data need to be exchanged across services without centralized control
- projects of data portability, especially in social networking
- integration / aggregation / interlinking of large collections of heterogeneous data (including non-structured and multi-media data)

- entity extraction, annotation and indexing of large collections of documents and multi-media assets
- knowledge management

To seize opportunities related to these scenarios, the OKKAM Foundation can be involved in several ways:

- a. as research partner on entity-centric reasoning, entity matching & search, etc.
- b. as data, service and technology provider for entity naming or identifier management systems
- c. as consultant on entity-centric applications

The expectation created so far is encouraging: In the last few months, the Consortium has been receiving several collaboration requests from third parties and the first pilot projects are starting to be configured.

## **5 Further Information**

Do not hesitate to visit <http://www.okkam.org/> or get in contact with the Project Coordinator:

**Paolo Bouquet:**

*Departement of Engineering and Information Science*

University of Trento

Via Sommarive, 14 - 38123 Trento (Italy)

email: bouquet@disi.unitn.it

Phone: +39.0461.882088

Mobile: +39.335.6006338