# OKKAM

# Enabling the Web of Entities. A scalable and sustainable solution for systematic and global identifier reuse in decentralized information environments.

**http://www.okkam.org/**


## Annual Report 2010

The OKKAM project has been a 30 months FP7-Integrated-Project, finished on June 2010. OKKAM aimed at enabling the Web of Entities by creating a scalable and sustainable infrastructure: the Entity Name System (ENS), which allows systematic reuse of global and unique entity identifiers in networked environments.

OKKAM ENS creates an "anchor", not authoritative IDs, to access information about the corresponding entity on the Web through extraction tools. Thus, any recognized entity is associated to an (OKKAM) ID that will be the same everywhere, independently of the original source, format, or the extractor used for recognition.

## Summary of Activities

During the last year of the project OKKAM project achieved its stated objectives. Effort during this period focused on two main objectives:

1. Consolidating and enhancing the R&D outputs generated during the project. A high-performance and clustered state-of-the-art prototype of the OKKAM Entity Name System (ENS 3.0) was produced. The resulting technology is not only a highly scalable and usable tool, but also a solid foundation for a whole generation of new applications and services.
2. "Going public". It comprised the creation of a community portal to foster community involvement, organization of ad hoc events, and focused demonstrations to interested communities.

## Important Work Areas

The Project carried out work in two key activities:

***Enhancing the OKKAM Entity Name System (ENS)***

The purpose of the OKKAM Entity Name System (ENS) is to manage entity identifiers and the foster of their global re-use. Most of the R&D efforts have been focused on providing a state-of-the-art and clustered high-performance architecture that supports the integrated prototype of the ENS. The core service of the OKKAM ENS is entity search, where storage, indexing and matching technology was built for finding an entity given its description. Resulting technology is transparent to users, not worrying about which resources or services they are accessing. The ENS comprised the development of the following core components:

- **Storage**: a scalable repository of entity profiles, in which billions of entities are assigned an ID and a profile, to distinguish one entity from another. Profiles are mainly name-value pairs, with no fixed schema, so they are open, flexible, and type-independent. Part of the information points to other known identifiers for the same entity, and links to authoritative web resources (if available) which describe the entity.

- **Matching**: requests from client applications arrive in the form of a bag of keywords or a collection of name value pairs (unstructured or semi-structured queries). The matching layer uses this information to send back to the application the ranked a list of the best candidates (if any) found in the ENS. Ideally, this is just one identifier, so we are striving for a very high Top-1 success rate;
- **Lifecycle Management**: it takes care of the evolution Storage of the repository and of all entity profiles through time. For example it manages the creation of new entities, updates their profiles based on log analysis, provides measures of popularity which can be used to improve ranking of results;
- **Access Control**: it makes sure that only authorized applications or users may use the available APIs, filters out mining queries (I.e. queries aimed at retrieving sets of entities based on a common property) and prevents malicious attacks.

***Bootstrapping the ENS***

The ENS was pulled up by refining the established OKKAMization process, in order to support populating the ENS and deploying the OKKAM-empowered tools suit as well as the application scenarios. Thus, the effort was mainly focused in these three areas:

- **Repository population**: the population of the entity repository, i.e. the number of entities managed in the ENS, has been a crucial factor for the acceptance of the ENS. We have harvested entities (together with an automatically created profile) from some popular public sources like Wikipedia/DBpedia, geonames, UNIProt. In the current production platform, a population about 7.5 Mio entities has been created in the OKKAM entity repository. Notwithstanding, it has been also shown in experiments that the ENS technology can scale up to at least 100 Mio. entities with only minor reduction in performance.

  OKKAMization is defined as the process of analyzing existing (structured and unstructured) content and content under creation, identifying entity references in this content (entity recognition) and equipping the entity references with the associated OKKAM IDs.

- **OKKAM-empowered tools suit**: 16 OKKAM empowered plug-in tools have been developed for fostering the adoption of the OKKAM infrastructure and easing the creation of OKKAMized content. These tools make it easy for users to annotate the content they publish with the identifiers coming from the ENS. Among deployed OKKAM-empowered tools are: MS Word, Outlook, Firefox, Internet Explorer, Foaf-o-Matic, Protégé, the NeOn ontology editor. These tools may enable the user to find the identifier for a named entity and store it with the content in the easiest possible way without getting familiar with new tools. The description of these tools is provided at -http://community.okkam.org.

- **Application Scenarios**: the following three entity centric applications were developed in close collaboration with the OKKAM application partners on top of the OKKAM ENS. Main objective was to showcase the benefits of the project, improve the reflection of the business requirements in the technical requirements of the OKKAM ENS and create blueprints for the adoption of the entity-centric approach. Also, the development of these application scenarios was performed based on results from a questionnaire that targeted better capturing of the business requirements and their alignment with the technical requirements.

  - **Enterprise Knowledge Management**: this application was created to find answers to SAP customers' problems and the experts in the SAP community portals, in a much faster and more precise way. The application identified relevant entities (and their relations) in a user's request and matched 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 could also be annotated with OKKAM identifiers.

  - **Authoring and Publishing**: we created an authoring environment based in exploiting the recognition of known entities in publishing text. In this way authors (specifically, scientists and journalists) could create and collect additional information faster and in a semi-automated way. For example using information about proteins in biological papers, past events in news items, etc.

  - **Semantic Search**: we built an entity-centric semantic mashup engine called Sigma (http://sig.ma/) on top of the Sindice search engine (http://www.sindice.com). Sigma can be used to send a query about a given entity (e.g. the European Commission) and will return not just a list of documents, but a structured profile of the requested entity..

## User Involvement, Promotion and Awareness

User involvement, promotion and awareness activities included the support and involvement through the OKKAM community portal, publishing the results of OKKAM research, bringing the findings to specific constituencies for validation through workshops, presentation of findings and incorporating project outputs in more general and scientific publications; available at http://project.okkam.org.

Activities focused in two main objectives:

- Taking the message to the whole community of stakeholders: 43 dissemination activities were performed during last year, comprising: 16 Dissemination/exploitation events targeting the commercial audience (including OKKAM presentation at the Web3.0 Venture Academy) and 27 dissemination activities targeting research oriented audience; including the organization of a tutorial at ESTC 2009 and presenting OKKAM results at international conferences (e.g. WWW2010, ESWC2010).
- Exploitation driven dissemination activities: efforts included agreeing first pilot projects for exploiting the OKKAM infrastructure (e.g. the joint OKKAM pilot projects with the Trento Tax Agencies and the city of Manor –Texas/USA-)

## Future Work or Exploitation Prospects

The Community Demonstrator “Entity-centric Approach” and the Exploitation-oriented demonstrator “OKKAM Results” consolidated the bases for an application-driven roadmap for future research, experimentation and deployment activities

OKKAM exploitation and commercialisation strategy was validated and adjusted with the help of investors, experts, other researchers and companies active in relevant domains. The results were incorporated into the OKKAM business plan, which included a dual approach: (1) the not-for-profit public part of the OKKAM infrastructure, which will be made available for the entire community of web users in the form of a public TRUST (the OKKAM Trust), under the control of a high-level internation board of protectors; (2) the business opportunities will be exploited through the creation of a start-up company, which will commercialize products and know-how developed in the OKKAM project. In a first phase, the start-up will act as Trustee of the OKKAM Trust."

## Further Information

- Project web site: http://project.okkam.org/ , which includes a collection of articles and reports have been published in the project website.
- Try and test our publicly available tools released during the OKKAM project (http://community.okkam.org):

| | |
|---|---|
| OKKAM Web Search: http://api.okkam.org/search | Allows search on the ENS repository through a web interface |
| Foaf-O-Matic: http://www.foaf-o-matic.org | ENS-enabled support to the management of FOAF profiles |
| Protègè Plugin: http://community.okkam.org/index.php/Downloads/Okkam-Enabled-Tools/okkam4p-protege.html | Extends the ontology editor Protégé with the possibility of using OKKAM Ids as URIs for ontology instances |
| NeON Plugin: http://community.okkam.org/index.php/Downloads/ENS-enabled-Tools/okkam4n.html | Same as above for the NeOn ontology editor |