OKKAM: Enabling a Web of Entities

Paolo Bouquet
University of Trento
Italy
bouquet@dit.unitn.it

Heiko Stoermer
University of Trento
Italy
stoermer@dit.unitn.it

Daniel Giacomuzzi
University of Trento
Italy
giacomuzzi@dit.unitn.it

ABSTRACT

Already in the early 14th century, the philosopher William of Ockham created the philosophical principle known as Ockham's Razor, which can be translated from Latin as "entities should not be multiplied beyond necessity". Today, one of the big - and yet unresolved - issues in integrating information distributed over a network of accessible data is the uncontrolled "multiplication" of identifiers which are used to refer to the same entity (e.g. a person, an event, an organization, a location, a product) across multiple heterogeneous sources. In this paper we propose a system called OKKAM that is currently under development to tackle this "Identity Crisis" on the Semantic Web; we discuss requirements, architecture, usage scenarios and services and experiments we have developed so far.

1. INTRODUCTION

In the W3C recommendation *Uniform Resource Identifier (URI): Generic Syntax* [1] , a resource is defined as "anything that has identity"[1]. This means that not only web accessible pages and documents are resources, but also people, cities and conferences; even the concept of "car" and the property of "being the owner" of a car are resources, which can be referred to and described as any other resource (e.g. in an ontology).

Despite the generality of the Semantic Web approach, here we want to suggest that – in practice – there is an essential difference between managing and reusing identifiers of resources which correspond to "things" (in a very broad sense, ranging from electronic documents to bound books, from people to cars, from conferences to unicorns) – we will call them *entities* –, and identifiers which correspond to abstract objects (like predicates, relations, assertions) – which we will call *logical resources*. Our thesis is the following: *while any attempt of "forcing" the use of the same URIs for* logical resources *is in principle likely to fail (as every application context has its own peculiarities, and people tend to have different views even about the same domain*[2]*), the same does not hold – or holds at a level which is philosophically interesting but of little practical relevance – for* entities. In other words, the claim is that there are compelling *theoretical reasons* why the Semantic Web (and any other semantically driven information system) should not force people to use shared URIs for logical resources, but only (or mostly) *practical reasons* why people do not use shared URIs for entities.

By analogy, our claim can be illustrated by considering the different difficulty of building white page and yellow page services. The former basically requires an efficient mechanism for listing entities, retrieving them, and distinguishing entities one from another; the latter always presupposes some taxonomy, which is typically either too general (and therefore does not help in discriminating services), or too specific (and therefore heavy to master for users), or too complex (not usable).

It should be clear that the problem of unique identifiers for resources (in its two flavors: logical resources and entities) is crucial for achieving semantic interoperability and efficient knowledge integration. However, it is also evident that 99% of the research effort is on the problem of (i) designing shared ontologies, or (ii) designing methods for aligning and integrating heterogeneous ontologies (with special focus on the T-Box part of the ontology). Perhaps because of its "practical" flavor, we must recognize that only a very limited effort has been devoted to addressing the issue of identity management for entities. For example, ontology editors, such as Protégé, support the "bad practice" of creating new URIs for any new instance created in an ontology. In our opinion, this problem is not only of general interest for the Semantic Web enterprise, but is one of the most critical gaps in an ideal pipeline from data to semantic representation: if we do not have a reliable (and incremental) method for supporting the reuse of URIs for the new entities that are annotated in new documents (or any other data source), we risk to produce an archipelago of "semantic islands" where conceptual knowledge may or may not be integrated, but ground knowledge is completely disconnected. And since the most valuable knowledge is typically about individuals, we take this to be an issue that should be attacked.

In this paper, we introduce the main requirements and

[1]'A resource can be anything that has identity. Familiar examples include an electronic document, an image, a source of information with a consistent purpose (e.g., "today's weather report for Los Angeles"), a service (e.g., an HTTP-to-SMS gateway), and a collection of other resources. A resource is not necessarily accessible via the Internet; e.g., human beings, corporations, and bound books in a library can also be resources. Likewise, abstract concepts can be resources, such as the operators and operands of a mathematical equation, the types of a relationship (e.g., "parent" or "employee"), or numeric values (e.g., zero, one, and infinity)' [1].

[2]This is what in [2], which was co-authored by one of the authors of this paper, was called the *distributed knowledge* argument.

a prototype implementation of OKKAM, a service for supporting transparent integration of knowledge about entities through simple identity management support. OKKAM can be described at two different levels:

- the basic services – which belong to a module called OKKAM-Core – provide APIs to create and store URIs for entities, to add/modify/remove informal descriptions of each entity, to index the resources in which knowledge about an entity is provided (e.g. ontologies, web pages);
- on top of OKKAM-Core, OKKAM offers a collection of advanced services, including searching for already existing entities (using different search criteria), extracting information about entities, ranking results, supporting the reuse of URIs for entities in ontology editing, and so on.

The structure of the paper is the following: in Sect. 2, we introduce our motivations and the resulting goals for our project in more detail. After that, in Sect. 3 we describe first basic services that have been implemented on top of OKKAM-Core, whereas Sect. 4 illustrates usage scenarios we have addressed with the system. Section 5 explains first experimental results. Finally, we conclude with a discussion of issues and an outlook on the further development of OKKAM in Sect. 6.

2. GOALS AND MOTIVATIONS

As soon as one starts thinking about the idea of an entity repository, the temptation of building what Craig Knoblock[3] called an EntityBase in one of his recent talks, is very strong. In short, an EntityBase can be thought of as an entity-centric knowledge base, where knowledge is organized around entities instead of schemas (e.g. relational schemas or even ontologies). In such an approach, any entity type would be characterized by a collection of attributes (for example, for entities of type book, some attributes can be “author”, “title”, “date_of_publication” “publisher”), whose semantics is known in advance and explicitly specified.

We called this a temptation, as it is extremely appealing (we would always know what we know about an entity), but also very dangerous, as it presupposes a commitment on the meaning of an attribute which cannot be guaranteed in most practical situation by a repository which aims at being open, extensible, global. Therefore, an important requirement for our service is that it is light and fast, which can't be confused with yet another attempt in the direction of CYC [4] or SUMO [5], as systems of this type offered useful approaches in certain areas, but have obviously not contributed to a solution of the identity problem in the (Semantic) Web. What we are aiming to provide is a *naming service* for entities and a *directory* profiles about these entities; we do not aim at providing a *knowledge base*, for the mentioned reasons.

An Entity Profile stores *untyped* data about entities which will support the human user or an application using the OKKAM API to process descriptions about entities, and in effect enable them to assess whether the entity they want to store knowledge about in their own local KB already has a URI in OKKAM, or whether they have to create a new one. We store untyped data for the reason that typing an entity's attributes would require us classify the entity, which would be in contrast with the abovementioned goal. We do not discriminate types of entities, because we explicitly want to be able to provide naming and descriptions for *any* entity.

Of course at first sight one could think about what types of entities would be described in OKKAM, such as persons, artifacts, locations, companies etc.; this could make it appear sensible to provide a basic set of typed attributes for these entities. But we envision the system also to provide support for less obvious applications such as Named Entity Recognition from the field of Natural Language processing, which we will talk about later, in Section 4. In these applications, entities might represent a location or a piece of text in a document, a document itself or a collection of documents, and we end up with an unlimited set of potential types of entries, which makes it impossible to provide a common set of typed attributes. Therefore it is our opinion that only untyped or even unstructured descriptive metadata can provide for the envisioned level of generality.

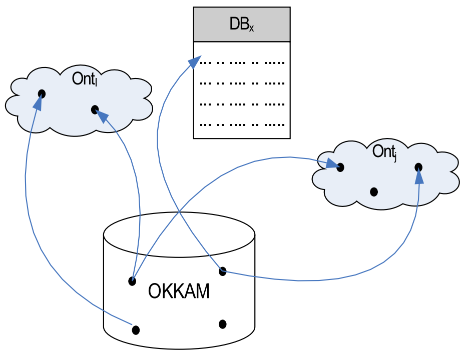


Figure 1: Schematic overview of OKKAM, plus external K/I sources

In addition to these untyped data, OKKAM provides for the management of what we call *ontology references* in the Entity Profile, i.e. a set of URIs to external sources that are known to store information or knowledge about these entities, as illustrated in Fig. 1. One of the reasons to go in this direction was the motivation to make OKKAM provide a possible solution to integration issues in the Semantic Web. While a great amount of work has been performed on schema-level information integration[4], the aspect of *entity-level information- and knowledge integration* still offers many opportunities for providing interesting approaches. One possible application we envision to support with OKKAM is an extension-based equivalence check for classes in an alignment or integration process. Currently, in

[3]Craig Knoblock's homepage: `http://www.isi.edu/~knoblock/`. Unfortunately, at this point no citeable publications about this topic are available. The facts mentioned here were presented at a seminar he gave at the ITC-IRST research center in Trento, Italy in 2006.

[4]It is hardly possible to cite all related work in this field. Specific to the area of the Semantic Web, the reader is referred e.g. to the publication list on `http://www.ontologymatching.org` for a host of publications, or the Ontology Alignment Evaluation Initative (`http://oaei.ontologymatching.org/`) which performs an alignment contest. For an overview more related to the database world, we refer e.g. to [7].

a schema-level integration process without extension check, classes can be *estimated* to be equivalent, but without an extension check the result of this estimation cannot be proved. Additionally, without a service that provides strong decision support about whether two individuals with the same name are actually identical or not (which is the current situation in the Semantic Web), an extension check will hardly deliver very reliable results. With the help of Entity Profiles in OKKAM we hope to improve this situation, because if we look at a case where two assumedly equivalent classes show that the sets of OKKAM-registered individuals associated to them are identical, we have very strong reason to support this equivalence assumption.

The last component of the Entity Profile is a set of assertions of identity between entities. We provide these for the case where two entities with different URIs in OKKAM are later discovered to describe the exact same object, and are thus identical. One possible criticism at this point is certainly the question how we can know and be certain about identity of entities. The answer is that we cannot. OKKAM will suffer from the same garbage-in-garbage-out property as any other information system. But with OKKAM at least we can provide a means for the (Semantic) Web to store and represent such information, and we hope that by consistent use of OKKAM in Web applications we can strongly improve the current situation by enabling agents to gain a certain level of confidence that they are actually “talking about” the same object.

3. OKKAM SERVICES

OKKAM can be viewed as a collection of services built on top of OKKAM-Core[5]. In this section we list the main services which – in our opinion – should belong to OKKAM, describe their implementation (if available), or present ideas on how they could be implemented.

3.1 Populating OKKAM

The first important service is the one that supports the population of OKKAM-Core with new entities. In fact, there are many issues that must be addressed and solved before new data is allowed to be stored. In particular, we want to stress the following:

- first of all, we want to add a new entity only if it is not already stored in OKKAM-Core. But this means that we need smart ways for recognizing if a new candidate entity is already stored, and for deriving when a new entity which looks like an entity already stored actually is a new one. These two requirements are crucial: failing to meet the first would lead to a lack of completeness (failing to support inferences which in theory are sound based on the fact that two names refer to the same entity); failing to meet the second would lead to a lack of correctness (false conclusions would be supported, based on the fact that two different entities have been collapsed onto a single identifier);

- imagine we detect that an entity is already stored, and that we find a new occurence of that entity in a document where some information about it is provided. Question: what if the new information conflicts with the old one? And, even before, how do we detect that there is such an inconsistency?

- as it will be clarified in the section on envisaged application scenarios, information may be imported in OKKAM-Core from very different sources, including humans (who may be carefully making data entry), ad-hoc wrappers designed to import entities from rich sources (e.g. lists of entities from Wikipedia), entity recognition tools (which may be extracting entity descriptions from free text). These potential sources may provide very uneven data, including a lot of garbage, which would undermine the role of OKKAM as a general and reliable tool;

- finally, a theoretical issue which needs to be addressed is the following: what does count as an entity? There is little doubt that people, organizations, cars, computer files, electronic devices, are entities. But, for example, is a document an entity? Is it an abstract entity, or it is identified with its physical realizations? If so, is every copy of a document a different entity? Another example is: are logical resources (like concepts, relations, topics) entities? Or the entity is the linguistic expression used to express a concept? But then are two linguistic formulations of the same concept different entities? And furthermore: are fictitious entities entities? Should we allow Pegasus and Spider Man to sneak into OKKAM-Core?[6]

To address these issues, we are developing the following compontents:

- *OkkamListsManager*

  On the WWW there are many lists of entities and are thus a potentially important resource for OKKAM. For example Wikipedia provides lists of countries, cities, members of particulars domains (e.g. Presidents of the United States, Computer Scientists, etc) that are exactly the types of entities that we want to store in the system. With the objective to find a standard mechanism for integrating these entities into Okkam we developed a language (an XML Schema) that describes the input that a data source has to follow to communicate with the population process of OKKAM. The main elements of the schema follows the internal structure of Okkam, in fact we have elements like ”Labels”, ”Label-prefix” or ”Label-value” that are easy to map with the tables elements of OkkamCore. This language is used by different wrappers that we developed and that try to convert the structure of a source list into the OKKAM input standard. For lists from the Web (Wikipedia, Yahoo, Google, etc.) the main purpose of the wrappers is the data cleansing process from HTML tags. After this step, the entity collection is normalized

[5]for a detailed description of the current OKKAM architecture and a discussion of other implementational aspects, please refer to [3].

[6]We notice that a very practical version of these philosophical questions is the following: what should be represented as an instance in an OWL ontology? And what as a class/property? The issue is tricky, and we make only one example: should “Pizza Margherita” be a class or an instance of an Italian food ontology? If we check e.g. [6], we find that the answer to this type of questions can be quite disappointing.

with the objective to delete duplicates. Entities with the same annotation label are recognized by the system and the OKKAM administrative user can check if there exist conflicts from members of the list that are the same entity (from a logical point of view). During insertion, for ach entity the system searches OKKAM if there is already an entity with the same label/s. If yes, this entity is "frozen" and included in a set of entities that should be checked by the administrator before addition to the system, otherwise it is added immediately.

- *OkkamDBManager*

  Another important information source for OKKAM can be generic databases, as far as we have access to them. Examples might include direct database access to information systems such as extranets, online shops or publishing houses. In this case the transformation from the internal structure of the tables into the OKKAM input language is easier because the main objective of the process is writing queries that build the link between the database structure and the okkam data structure. When the transformation into the input language is completed, the rows that come from the database follow the same process that we already describe with web lists. With database sources the role of the user becomes more important because, with high numbers of entities, duplication and redundancy are an increasing problem.

- *OkkamManualEntry*

  Another solution we provide to insert new entities is the manual case. A Web interface provides easy access to the insert function. The user can add new entities, with labels, ontology references, etc., to the system using a form to specify all the information that he/she want describe the new entity with. As in the previous case, if the system finds a possible conflict with entities that are already in Okkam, it issues a warning message that informs the user of the possible error. This methodology of insertion is the slowest that Okkam provides, but it is the most precise and complete because the user can provide information that the system can not automatically discover, and optimize the input in a feedback loop.

- *Protégé Plugin*

  We provide a plugin for the ontology editor Protégé which we describe in further detail in Sect. 4.1.

3.2 Searching for URIs

Another critical service – which is currently under development – is searching for the identifier of an entity which is known either by description (e.g. the name in case of a person), or by an identifier which was not issued by OKKAM. This web-based service should be held very simple, like a traditional search engine, and based on an easy mechanism to visualize results. In a standard use case the user types in a keyword associated to an entity and the system searches the repository for instances that match this label. The returned data will be a representation of the data in the entity profile (URI of the entity, the other labels associated to the entity, and the classes of the ontologies in which the entity is used).

This envisioned Web site of OKKAM is not the only application exploiting search functionality. There are situations in which it is very unlikely that users will consult the OKKAM web site to search the URI of the resources that they need. For example, if we have a large database with all employees of an organization is impossible that the designers and developers wanting to build semantic application on this data search in the OKKAM web site all the URI's of the persons stored in their database. This process can be simplified if they can use an automatic service, in this case a web service, that provide an access point to OKKAM that an application can use. The developers can build an application that extract the data from their database and send them to the web service which will return some results, URI, about the information that already are stored in OKKAM.

4. USAGE SCENARIOS

4.1 Runtime support for ontology editing

Another important area for which OKKAM has to provide services and applications are existing Semantic Web tools. In particular, ontology editors are applications where users build a formalization of part of the world by means of classes and instances of these classes, all identified by URI's. One of the most widely used and important editors is Protégé, an open source product that can be extend and modified with "plug-ins" added on the core system. For the OKKAM vision it is of high importance to develop a plug-in for this application which provides a connection with the URI database when users create new instances of a class, which we are doing as illustrated in Fig. 3. If a user creates a new instance of a class, instead of assigning an arbitrary, meaningless number as ID the plug-in will search the repository whether an URI already exists that can be assigned to this new instance. The selection process is envisioned similar to the web search use case where a list of URI's that match the label for the new instance are visualized to the user.

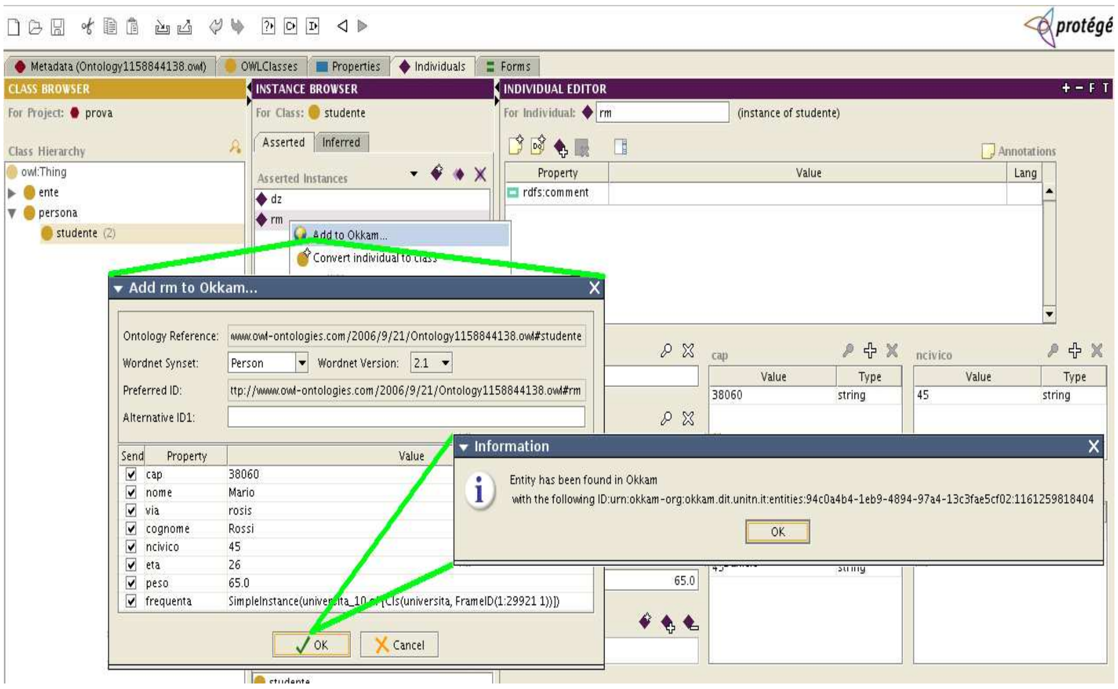
Important support for all the selection processes comes from additional tools, as for example WordNet, that provide information about the meaning of the classes used in the ontologies where the new instances are created. With this information the system has more data to try to recognize the correct URI to return to the users or application that query OKKAM.

4.2 Supporting Knowledge Extraction and Representation

One of the scenarios we are currently implementing with the help of OKKAM is to support Knowledge Extraction (KE) processes and the resulting Knowledge Representation (KR) in a Semantic Web project[7] that aims at building a large-scale Knowledge Base (KB) from information stored in distributed document bases. The architecture comprises a pipeline of processes that covers all steps from KE to the building of the KB (the so-called *Semantic Resource Network*) for end-user services, as illustrated in Fig. 3 and described in detail in [8].

Within the pipeline there are several points of application imaginable, two of which we have currently implemented

[7]see `http://www.vikef.net` for further information about the VIKEF project.



Add rm to Okkam...

Ontology Reference:	www.owl-ontologies.com/2006/9/21/Ontology1158844138.owl#studente
Wordnet Synset:	Person
Wordnet Version:	2.1
Preferred ID:	ttp://www.owl-ontologies.com/2006/9/21/Ontology1158844138.owl#rm
Alternative ID1:	

Send	Property	Value
✓	cap	38060
✓	nome	Mario
✓	via	rosis
✓	cognome	Rossi
✓	ncivico	45
✓	eta	26
✓	peso	65.0
✓	frequenta	SimpleInstance(universita_10 of [Cls(universita, FrameID(1:29921 1))])

cap

Value	Type
38060	string

ncivico

Value	Type
45	string

Information

Entity has been found in Okkam
with the following ID:urn:okkam-org:okkam.dit.unitn.it:entities:94c0a4b4-1eb9-4894-97a4-13c3fae5cf02:1161259818404

Figure 2: A Protégé plugin for generating individuals registered in OKKAM.

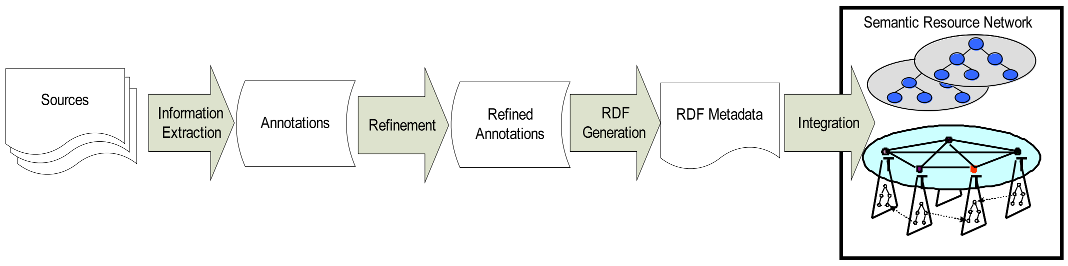


Figure 3: Knowledge pipeline to be supported by OKKAM

and are further described in the following:

- *Information Extraction: Named Entity Recognition and Coreference*

  Whenever the NLP process recognizes a named entity in a piece of text, it interacts with OKKAM to analyze whether this named entity already has a unique URI. If yes, the NLP process stores locally[8] the fact that a uniquely identified entity has been discovered with additional information such as its location in the document, etc. If the entity does not have a URI yet, an Entity Profile is created in OKKAM and the resulting URI is used accordingly. For subsequent discoveries of the same named entitiy, the same URI will be used to indicate that the two discovered entities are in fact the same, just in different locations of the document. This approach is equally applicable to discovered coreferences[9]. If the NLP process updates the Entity Profiles in OKKAM correctly, we gain direct access to search situations of the type "show me all documents that talk about this entity", as the respective links would be stored as Ontology References which we can evaluate and reason about with a higher-level service.

- *Refinement: Identity Discovery*

  In the refinement phase, as depicted in Fig. 3, we can address shortcomings of the NLP processes in terms of discovery of identity. The VIKEF pipeline has dedicated a whole processing step to this issue, as – at the named entitiy extraction level – it is not always possible to detect identity between entities. Obvious examples in this case are missing correspondences between orthographic variations hinted at already in Sect. 3.2, e.g. the fact that within one document there is a certain probability that the strings "Stoermer", "H. Stoermer" and "Heiko Stoermer" denote the exact same individual. With support of the OKKAM system, we have implemented several heuristics to address this issue, the simplest performing a substring query to OKKAM and using a string similarity measure on the results to choose candidates for establishing an assertion of identity between them, and thus to cluster annotations. A higher level process is free to either choose one single URI for all the annotated entities or to retain the original URIs, as it is always possible to perform clustering via analysis of identity assertions in OKKAM.

4.3 Entity-based Information Integration

Even though the problem of identifiers arises both for abstract objects (e.g. classes of objects, relationships between classes of objects, etc.) and for concrete individuals (e.g. Plato, the University of Trento, Paris, the WWW2007 Conference), the problem seems to be conceptually different for abstract objects and concrete individuals, and therefore its solution may require different methods and tools.

Establishing whether two elements belonging to different information sources (e.g. attributes in two different relational databases, classes in two different ontologies) provide information on the same abstract object (if the two attributes express the same property, or if the two classes are logically equivalent) is not simply a matter of identifiers, but also of sophisticated reasoning on the intended meaning of the two elements in the two sources, and this may presuppose the use of a large amount of background knowledge, of lexical repositories, of techniques for instance-level comparison, and so on. In fact, the heterogeneity in conceptually modeling some domain of discourse seem to be so deeply rooted in our human nature, that several researchers and practitioners seem to agree on the conclusion that proving an identity across abstract objects is extremely hard; and that working on semantic mapping between abstract objects is a long-standing task for the scientific community.

The same does not seem to hold for concrete individuals. In fact, in many real life domains (e.g. in identifying people living in a country, computers and mobile devices on a network, published material, and so on), uniquely identifiers for individuals have been introduced since a long time, and the adopted solutions does not seem to require any sophisticated theory. However, if we consider the universal space of information offered by the Web (and any other large-scale open network-based application), there are two major issues which need to be addressed:

1. there is no global system for generating (and storing) public unique identifiers for individuals which are mentioned in digital documents. Of course, there are global identifiers for special types of concrete objects (e.g. URLs for web pages, DOI for digital documents, and so on), but nothing like that is happening for the concrete individuals mentioned in digital documents. And we don't mean only people, but also locations, events, organizations, and so on.

2. even when systems for managing identifiers are provided, their purpose is typically not to make public identifiers globally available for reuse in any new type of digital document. Some systems are meant to work only on limited domains, or for a restricted type of users. Others aim at supporting people in authenticating themselves when logging into different applications, but identifiers are not supposed to be used to refer to an individual in a generic document; other systems provide IDs for special types of objects which are used for classification purposes. But, if one needs to mention the University of Trento in a text document, in anontology or in an HTML page, there is no obvious location where a public (and universal) ID for it can be found, besides the URL of its web page (which is not the University itself).

As an exemplary case, consider the Semantic Web domain. The W3C has provided a recommendation for a syntax to build universal identifiers (URIs) for arbitrary resources. However, there is no systematic support to find and possibly re-use URIs previously introduced for a given resource in the universal information space of the Web; and thus a new universal (but local) URI is created every time an entity

[8]In fact, the annotations created in this phase are stored in an XML file, which is later refined and then used as a base for the generation of RDF annotations that will be fed into a large knowledge base.

[9]A coreference is a linguistic pattern typically involving pronouns when talking about an object that has previously been named. Example: "Peter is a good runner. He does 10k in 45 minutes." The personal pronoun *he* establishes the coreference in this case.

is mentioned in some RDF/OWL knowledge base. And indeed all ontology editors - including the well-known Protégé editor from Stanford - do nothing to prevent this "bad practice"; and the major Web and Semantic Web conferences (WWW, ESWC, ISWC) produce RDF graphs where the same concrete individuals are referred to via different URIs, which makes their data-level integration not so smooth.

5. EXPERIMENTS AND RESULTS

To support our argument regarding information integration as mentioned in Sect. 4.3, we performend a preliminary experiment based on RDF metadata about ISWC2006[10] and ESWC2006[11] metadata sets.

Upon analysis of the RDF data it becomes evident that a plain RDF-style integration of the two graphs would not produce very usable output because the naming and identification schemes for individuals differ, which supports our general claim in favour of a system like OKKAM.

Our goal was to align the sets of *person* entities described in the RDF data, by exchanging their original IDs with identifiers from OKKAM ("okkamization"), and thus hopefully reaching a point where the statements of the two data sources could be fed into one knowledge base, e.g. to answer queries such as "who attended both conferences" or "who published a paper at both conferences", which would be impossible without this form of alignment on the instance level.

To this end, we have developed a software module that okkamizes the first data source by inserting an entity profile for every extracted Person entity into OKKAM, and replacing the existing identifier with the resulting OKKAM-ID. Based on this population[12], we are able to perform a more refined okkamization of the second data source, as for every Person entity we query OKKAM for the existence of this entity (applying a set of heuristics to tackle some of the issues described in Sect. 3.2), and – if the entity already has a profile that matches – we re-use the OKKAM identifier. Resulting are two data sets that use the exact same identifiers for entities that the system believes to describe the same object.

Table 1 describes the results of the experiment. Out of the complete number of individuals in both files, we extracted the entities that are an (inferred) individual of `foaf:Person`[13]. After the okkamization process, OKKAM detects 27 entities that appear in both data sources. A manual check of both sources reveals that we achieved a recall and precision of 100%.

	ESWC	ISWC
Number of individuals	615	579
Number of Person entities	189	316
Okkam matched Persons	27	
Manual match	27	
Precision & Recall	100%	

Table 1: Integration of ISWC and ESWC *Person* entities

Of course, these results cannot be taken as a performance claim for our system. They have to be regarded as a first experiment in a very restricted and defined test bed: we knew which types of entities we wanted to match, and we had heuristics implemented in software that performed more precise matching than only a string match on the ID. We used this first experiment as a proof-of-concept, also to establish and run the architecture vertically, from data layer through services to an automated client application. In the next implementation phases of OKKAM it is obvious that new algorithms for search and matching have to be implemented on a far more general level.

6. DISCUSSION AND CONCLUSION

OKKAM is the typical example of a system which is not based on some radically new scientific result, but aims at filling a gap by using existing technologies in a new way. In our opinion, without OKKAM (or a similar service) most Semantic Web promises will never be kept, as it provides a sort of bottom level for integration which cannot be achieved *ex post* when the ball stops. However, the fact that the basic technologies are already available should not lead us to underestimate the critical factors which may affect the success and adoption of OKKAM. In addition to aspects already discussed throughout the paper, we identify acceptance issues in the form that not every party involved in the Semantic Web may be willing to use a centrally managed service that is outside of their control. Privacy issues include all the well-known aspects of data security, access management, privacy etc. that almost all public information systems share. Last, but not least there are of course questions of offered features and functionality, such as a really efficient and intelligend search and ranking mechanism for Entity Profiles in OKKAM, as well as performance and scalability issues which are again common to most information systems. Our planned next steps are to address exactly these issues in the form of further research and by developing additional services on top of OKKAM-Core.

We conclude with the statement that currently, when creating ontologies, people actually perform two different tasks: they specify a conceptualization, and then "populate" such a conceptualization with instances by assigning instances to a class and specifying the values for properties (if any). It is a trivial observation that the same domain (set of entities)

[10]The ISWC2006 data were retrieved in January 2007 through a SPARQL graph construction query on the SPARQL endpoint provided under the address `http://128.192.251.191:8080/joseki/iswc` as indicated on the ISWC2006 website (`http://iswc2006.semanticweb.org/program/tech_links.php#core`). At the time of writing of this paper, this address is unreachable; for this reason we cached a copy of the resulting graph in RDF/XML serialization under the following address: `http://www.okkam.org/publications-1/supportfiles/iswc-eswc-experiment/iswc.rdf/view`

[11]the ESWC2006 dataset is provided for download as a set of individual RDF files from `http://www.eswc2006.org/rdf/eswc2006-rdf-descriptions.zip`. These files are only partially valid and cannot be integrated automatically without the resulting document being error-free; thus we performed some manual corrections to the input to at least reach a point where a combined single file would parse. The resulting integrated file is provided under the address `http://www.okkam.org/publications-1/supportfiles/iswc-eswc-experiment/eswc_joined.rdf/view`

[12]the experiment started from a completely empty installation of OKKAM to have no interference with existing data

[13]`http://xmlns.com/foaf/0.1/Person` is the class used by both RDF graphs to denote a person entity.

may be used to populate different ontologies (e.g. we may have two different conceptualizations of Italian wines&food), and that any two ontologies (e.g. an ontology about semantic web researchers and another about people living in Italy) may have overlapping domains. Creating a conceptual schema and then populating it with instances address two different issues: the first is an *epistemological* issue (it has to do with knowledge about the world), the second is an *ontological* issue (it has to do with existence).

From a design perspective, what we propose is to keep these two tasks separated: on the one hand, we need a universal and non ambiguous way to refer to the entities about which an agent may have some knowledge; on the other hand, we need a way to specify knowledge about these entities. We believe that the help of OKKAM this goal can be achieved more cleanly for the Semantic Web, as to existing methods of specifying knowledge in the form of ontologies and knowledge bases we add an identity and reference architecture with a central character that enables systems and agents to ensure that they "talk" and store knowledge about the same entities, if these objects share the same identifier.

7. ACKNOWLEDGMENTS

This research was partially funded by the European Commission under the 6^{th} Framework Programme IST Integrated Project VIKEF - Virtual Information and Knowledge Environment Framework (Contract no. 507173, Priority 2.3.1.7 Semantic-based Knowledge Systems; more information at `http://www.vikef.net`)

8. REFERENCES

[1] T. Berners-Lee, R. Fielding, and L. Masinter. *RFC 3986: Uniform Resource Identifier (URI): Generic Syntax.* IETF (Internet Engineering Task Force), 2005. http://www.ietf.org/rfc/rfc3986.txt.

[2] Matteo Bonifacio, Paolo Bouquet, and Paolo Traverso. Enabling distributed knowledge management: Managerial and technological implications. *Informatik - Zeitschrift der schweizerischen Informatikorganisationen*, 1:23–29, 2002.

[3] Paolo Bouquet, Heiko Stoermer, Michele Mancioppi, and Daniel Giacomuzzi. OkkaM: Towards a Solution to the "Identity Crisis" on the Semantic Web. In *Proceedings of SWAP 2006, the 3rd Italian Semantic Web Workshop, Pisa, Italy, December 18-20, 2006. CEUR Workshop Proceedings, ISSN 1613-0073, online http://ceur-ws.org/Vol-201/33.pdf*, December 2006.

[4] Douglas B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):32–38, 1995.

[5] Ian Niles and Adam Pease. Towards a standard upper ontology. In *FOIS*, pages 2–9, 2001.

[6] Natalya F. Noy and Deborah L. McGuinness. *Ontology Development 101: A Guide to Creating Your First Ontology.* Stanford University. `http://protege.stanford.edu/publications/ontology_development/ontology101.html`.

[7] Erhard Rahm and Philip A. Bernstein. A Survey of Approaches to Automatic Schema Matching. *VLDB Journal: Very Large Data Bases*, 10(4):334–350, 2001.

[8] Rodolfo Stecher, Claudia Niedere, Paolo Bouquet, Thierry Jacquin, Salah At-Mokhtar, Simonetta Montemagni, Roberto Brunelli, and George Demetriou. Enabling a Knowledge Supply Chain: From Content Resources to Ontologies. In *In Proceedings of the ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation, Budva, Montenegro, June 12, 2006. CEUR Workshop Proceedings, ISSN 1613-0073, online http://ceur-ws.org/Vol-187/16.pdf*, 2006.