UNIVERSITY
OF TRENTO

DIPARTIMENTO DI INGEGNERIA E SCIENZA DELL'INFORMAZIONE

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.disi.unitn.it

A COGNITIVE CONTRIBUTION TO ENTITY REPRESENTATION AND MATCHING

Barbara Bazzanella, Paolo Bouquet, and Heiko Stoermer

January 2009

Technical Report # DISI-09-004

A Cognitive Contribution to Entity Representation and Matching

Barbara Bazzanella [a] , Paolo Bouquet [b] , Heiko Stoermer [b]

[a] *University of Trento, Department of Cognitive and Education Sciences, Via Matteo del Ben, 5, 38068 Rovereto, Italy*
[b] *University of Trento, Department of Information Science and Engineering, Via Sommarive, 14, 38100 Trento, Italy*

Abstract

The problem of Data Linkage in the Semantic Web can be divided in two lines of action: schema and ontology matching/mapping, which allows us to draw conclusions about sets of individuals through concept relations, and entity-level linkage, where more information can be reached from distributed sources because of the fact that the information is about the same entity. While the area of schema and ontology matching is traditionally much addressed, it appears that today the Semantic Web looks very much like a collection of "information islands" that are very poorly integrated with each other, especially on the individual level; and when some of these islands are linked, this is often the result of a lot of hard and time-consuming manual work. The general problem we are working on is to provide a structured approach of how to improve the situation of data linkage at the level of individuals in the Web of Data. As a specific contribution, in this article we describe an empirical investigation about how humans describe individuals (or *entities*), by analyzing a feature-listing experiment performed by a large sample of participants. We propose a measure of relevance to analyze the results, and apply the findings to the specific problem of entity matching in a large entity repository, by proposing a novel approach for entity matching/alignment. We show in a first experimental evaluation that such an approach, which takes into account the cognitive point of view of entity representation by humans, can provide an improvement over other relevant approaches.

Key words: Identity, Reference, Entity Representation, Entity Matching, Semantic Web, Web of Data

1. Introduction

In a very early note published in 1998 [1] , Tim Berners-Lee describes his vision of the Semantic Web as a global space for the seamless integration of countless semantic knowledge bases into an open, decentralized and scalable knowledge space. Much progress has been made since then to make this vision happen, but we must note that the efforts have not yet made this vision a reality. It turns out that one of the main reasons seems to be that today the Semantic Web looks very much like a collection of "information islands" that are very poorly integrated with each other; and when some of these islands are linked, this is often the result of a lot of hard and time-consuming manual work.

Ideally, the integration of information islands into a global Semantic Web should be based on the practice of using a URI for referring to any type of resource in RDF/OWL content. The key concept is that "[t]he global scope of URIs promotes large-scale *network effects*: the value of an identifier increases the more it is used consistently" [2] .

Email addresses: b.bazzanella@email.unitn.it (Barbara Bazzanella), bouquet@disi.unitn.it (Paolo Bouquet), stoermer@disi.unitn.it (Heiko Stoermer).

[1] See http://www.w3.org/DesignIssues/RDFnot.html.

[2] See *Architecture of the World Wide Web, Volume One* (W3C Recommendation 15 December 2004) at http://www.w3.org/TR/2004/REC-webarch-20041215/.

When referring to resources, we face the "Semantic Web version" of two well-known problems in information integration:

- heterogeneity of vocabulary: the same concept (e.g. "person") or property (e.g. "first name") may be referred to through different URIs, and therefore may not be recognized as the same concept or property in two different vocabularies;
- entity recognition: the same real world object (e.g. "Florence") may be assigned different URIs in different RDF repositories, and therefore may not be recognized as the same entity.

While the first issue is widely recognized and investigated [3], for a long time the second was largely neglected in the Semantic Web community, though it received – and is receving again – a lot of attention in the database community (under the headings of record linkage, data deduplication, entity resolution, etc. [7,4]).

However, this concentration on schema issues looks like a serious strategic mistake, because it only addresses part of the integration problem, while the other part is neglected.

2. Background and Problem Statement

To address the aforementioned problem, the EU-funded OKKAM project deals with the creation of the so-called Entity Name System (ENS) [2], an open, public back-bone infrastructure for the (Semantic) Web that enables the creation and systematic re-use of unique identifiers for entities. This is achieved by implementing a large-scale infrastructural component and services for describing entities, and assigning identifiers to them, so that other users can benefit from a network effect and re-use identifiers for entities that have already been described in their own information systems.

The ENS can be thought of as a very large, distributed "phonebook for everything" [4]. Users and systems using identifiers issued by the ENS benefit from the fact that this a-priory convergence on identifiers for entities leads to a high integrateability of information, and – in Semantic Web terms – enables correct graph merging, and thus a real global knowledge space, without the need for ex-post deduplication or entity consolidation.

The definition of "entity" in the ENS is purposely given in a very broad fashion, and covers all kinds of things from "anything that an information system talks about" to "an individual in an ontology" or "the interpretation of a variable in a first-order theory". The reason for this very un-precise approach is the simple fact that – even though the creators of the idea have a sort of wishful thinking regarding the types of objects that should be covered – in reality it will be impossible to predict what finally enters into the system once it opens to the public.

The consequence is that in order to describe such entities in the ENS, it was decided to *not* impose or enforce a certain schema to be used for the description of different types of entities, as well as strong typing of entities is not pursued or enforced. In fact, it is possible to create completely free-form, key/value based descriptions for entities (similar to "tagging" in folksonomies), which allows for complete genericity, without the need for a commitment to a certain formalism that would anyway be disputable, or to very abstract top-level categorizations as we know them from the area of upper-level ontologies.

However, such genericity obviously has its downsides: the ENS can never *know* what type of entity it is dealing with, and how the entity is described, due to an absence of a formal model. This becomes very relevant when searching for an entity, a process that we call *entity matching*. The envisioned use of the ENS is that an agent (human or artificial) has a certain entity in mind and provides a description of this entity, which is then used for finding and reusing the entity identifier, similar to the use of a traditional search engine to find the desired target of an HTML hyperlink [5]. The state of the art in related fields that can provide solutions to answering such matching queries has been analyzed in [24], and we come to the conclusion that the "no free lunch theory" of Wolpert and McReady [6] holds also for entity matching: to achieve a performance (in our case: precision) that is better than what a generic algorithm can provide, a *set of specialized algorithms* is required.

[3] See e.g. [5] for a recent survey of approaches and tools for schema-level alignment of ontologies.

[4] As opposed to a "knowledge base about everything", which is by definition *not* the aim of the ENS.

[5] This is in contrast with mining queries which we do *not* support. Such queries would have as a goal to find a set of results that fulfil certain criteria, such as "all single males between 30 and 40 years that live in a high-income neighbourhood".

[6] See `http://www.no-free-lunch.org/` for a rich source of information.

Such specialized algorithms can be adapted from fields such as name matching, address matching, named entity recognition, geo information systems, schema matching, approximate joins in databases, etc., but at this point the ENS faces two problems: first, as mentioned before, we cannot assume to know what kind of entity we are dealing with, and secondly, we cannot rely on homogeneous descriptions of entities (i.e. even if we knew the type of entity, we cannot assume that two entities of the same type are described using the same schema).

3. Objectives

To resolve this conflict between generality and precision, we attempt to foster the convergence of entity descriptions on a small set of default types, and attributes for these types, by providing *suggestions*: when a new entity is to be created in the ENS, an agent has the possibility to select a default type and description, and "fill in the blanks", or otherwise to provide any other kind of description. With this approach we hope to achieve useful clustering in the ENS, which puts us in a better position for entity matching, because at least in some cases we can understand better what kind of entity is described, and how it is described, which allows for a far better development and selection of specialized matching algorithms.

The first part of the work we are describing here is an experiment that has been performed to establish the mentioned suggestions for entity types and their descriptions. Instead of simply accepting (or inventing out of our mind) a certain schema, we decided to use a bottom-up approach of schema creation, and thus asked – with the help of a public poll – several hundred participants from different linguistic backgrounds how they would describe certain types of entities. The results presented in this article are significant from several points of view. First of all, it fulfills our need for "real-world" default schemata for which we can expect high acceptance in our context. Secondly, to the best of our knowledge, an experiment of this type has not been performed so far, and thus contributes to the state of the art in knowledge representation through ontologies. Thirdly, it allows for interesting reverse conclusions from values to entity types, e.g. we can infer an entity type from a descriptive value (see also Sect. 4.3). And finally, it provides very detailed insight which attributes are relevant for the description of certain entities, which can help knowledge engineers evaluate and improve their systems. Due to the fact that the results of the analysis are very extensive (beyond 100 pages), we also refer the reader to an acompanying technical report that we have published, which contains all details [1].

In the second part, we underline the usefulness of the data we gathered and analized, and present a novel approach for entity matching which bases on the insights we have gathered in the mentioned experiment. The approach bases on a new similarity score that takes into account not only the similarity of features, but also the circumstance that certain features are more meaningful for identifying an entity than others. We have implemented a prototype, created an experimental data set and a benchmark, and compared the resulting algorithm to other relevant approaches.

The rest of the article is structured as follows: on the methodology side, Sect. 4.1.1 explains the rationale of how we selected a set of top-level categories for entities in the ENS, and Sect. 4.1.2 describes in detail the experiment we have performed. Sect. 4.2 presents an excerpt of the extensive results of the experiment, together with the statistical measures used to perform ranking of the results. A high-level description of several novel ideas of how to apply the findings to the area of entity representation and matching is given in Sect. 4.3. Section 5 describes the proof-of-concept implementation of an approach that applies the findings to the area of entity matching, including formal definitions and experimental evaluation. Section 6 illustrates ongoing and future work that is in preparation, and Sect. 7 concludes the article.

4. The Entity Identification Experiment

4.1. *Methodology*

4.1.1. *Selecting Top-level Categories*

The first step in the entity representation experiment we are presenting here was to select an appropriate collection of top-level categories (classes, concepts) which we can suggest to users for a "weak" or "light-weight" classification for the entities they create. We identified four main requirements for this collection:

Usefulness. The set of top-level categories needs to be useful for a "normal" user, in that the concepts cannot be too abstract or too specific.

Disjointness. The categories need to be selected in a way that makes it easy to decide whether an entity belongs in one or the other, optimally through disjointness of the categories.

Conciseness. The number of categories should stay within easily manageable bounds, optimally below the "magic" number of 7 items [14], so that a user can decide at a single glance without further investigation which category should be chosen.

Coverage. The set of categories should be made in a way that all the entities that we envision to enter into the "population" of the ENS can be assigned to one of the categories.

In order to achieve these goals, we adopted a top-down approach: we analyzed the main top-level ontologies available in literature (Wordnet [6], Dolce [17,11], Sumo [15] and Cyc [12]), to integrate important ontological distinctions from those ontologies. At the end of our analysis we identified the following six top-level categories [7]:

- PERSON
- ORGANIZATION
- EVENT
- ARTIFACT
- LOCATION
- OTHER

We point out that the last category (OTHER) is a miscellaneous category that contains all entities that are not classifiable in one of the other categories and formally can be thought of as the complement of the union of the first five categories.

Another aspect that should to be mentioned is the level of abstraction of our categories. Our choice was guided by two constraints. The first is related to the cognitive reliability of the categories: it is well-known that categories are organized into a hierarchy from the most general to the most specific, but the level that is cognitively most basic is in the *middle* of the hierarchy [20]. The second is more connected to the assumed use of the final system. Directly connected to the latter constrain is the choice about the first category, PERSON. Although a more general category, such as Being, would allow us a better ontological coverage, including for example animals, it is not very probable that this latter type of entities would populate our system in large numbers.

As evident from the list, we limited our analysis to a subclass of entities that we can describe as "physical" entities (things that have a position in space and/or time), missing out "abstract" entities (things that do not have spatial nor temporal qualities, and that are not qualities themselves). The distinction between physical and abstract entities is at the base of the SUMO ontology (physical entity vs. abstract entity), the DOLCE ontology (endurant, perdurant particular vs. quality and abstract particular) and the CYC ontology (Intangible thing vs. Individual thing). The notion of abstraction is also present in WordNet, but has a different ontological coverage, not referring to state, psychological feature, action and phenomenon.

Following the distinction proposed by the CYC Ontology, we can distinguish between temporal entities and spatial entities, which justifies two of our top categories: EVENT and LOCATION. An EVENT is a thing that occupies a point (or period) in time, whereas a LOCATION is a thing that occupies a space. Both can have spatial and temporal parts, but the ontological nature is determined only by the *essential* parts that are temporal for events and spatial for locations.

Another important ontological assumption that we followed to build our list of top-level categories is related to the behavior of the entity in time. This distinction is connected to the difference between what philosophers usually call "continuants" and "occurrents", or using the terminology adopted in the Dolce framework between "endurants" and "perdurants". The main idea is that there are entities (endurants) that are wholly present (all their parts are present) at any time at which they exist and other entities (perdurants) that extend in time and are only partially present for any time at which they exist because some of their temporal parts may be not present. This motivated us to distinguish between entities that *are* in time like for example PERSON or ARTIFACT and entities that *happen* in time like EVENT, keeping another distinction that we can find both in the Sumo ontology (object vs process) and in the Dolce ontology (perdurant vs endurant).

A further ontological distinction we made within our basic categories is related to "agentivity". This property refers to the attribution of intentions, desires and believes and the ability to act on those intentions, desires and believes. On the basis of this assumption we can distinguish physical entities that

[7] We use small caps notation for the list because we want to denote the category itself, and not a natural-language label for the category. We could have chosen to use single characters as for variables, but decided to use this kind of notation of easier readability.

are agentive such as PERSON (or groups of several agents operating together like ORGANIZATION), and entities that are not-agentive such as ARTIFACT.

Another difference that is taken into account is that between "Individual" entities and "Collection". This ontological constrain is evident both in Sumo and CYC, and is used to explain the notion of collective entities such as ORGANIZATION, whose members can be added and subtracted without thereby changing the identity of the collective.

Similarly, WordNet distinguishes Entity (defined as something having concrete existence, living or non-living) and Group (which is any number of entities considered as a unit).

After making explicit the representation of the so-called ontological commitments (abstract vs physical, temporal vs spatial, endurant vs perdurant, agentive vs non-agentive, individual vs collective), we can provide definitions of each of our top-level categories.

Definition 1 (Person) *A physical entity, endowed with temporal parts that can change as a unit (endurant) and able to express desires, intentions and believes (agent).*

Definition 2 (Organization) *A physical collective entity, whose members are intelligent agents. In terms of behavior in time, an organization changes in time as a whole object so we can define it an endurant. As a collection of agents that operate together, an organization can be considered an agentive entity, characterized from desires, intentions and believes.*

Definition 3 (Event) *A physical individual entity that happens in time, perdurant.*

Definition 4 (Artifact) *A physical entity intentionally created by an agent (or a group of agents working together) to serve some purpose or perform some function. An artifact is a non-agentive endurant.*

Definition 5 (Location) *A physical individual entity that has a spatial extent, endurant.*

Definition 6 (Other) *Any entity that cannot be categorized in any of the above categories.*

4.1.2. *The Experiment: Approach and Implementation*

After establishing the top level categories of our study, we conducted an experiment in order to evaluate how people describe entities belonging to such categories.

The goal of this experiment was investigating which attributes are considered more relevant by people to identify types of entities selected as exemplars of the main categories reported above.

In order to get subjects to generate a representation of the categories investigated, and to derive from it a set of representative attributes, we adopted the feature-listing task paradigm [8]. In a typical feature-listing task, participants are presented with a set of category names and are asked to produce the attributes they think are important for each category. We adapted this paradigm to our purposes, inducing subjects (through scenarios) to produce lists of attributes they think not generically important for each category but relevant to identify uniquely members of the category. Consequently a participant's list of attributes is assumed represent a sort of temporary abstraction (not exhaustive of the complete knowledge that the subject has about the category) that contains the main attributes relevant for the specific task. Because of the dynamic nature of feature listing results and the linguistic nature of the task, we expected a certain variability both across and within participants. To deal with this variability we tested numerous subjects (N=358) and then pooled responses to detect a subset of attributes used systematically by participants. Finally we quantified the importance of each attribute by means of a single averaged measure of relevance.

4.1.2.1. *Approach* Since our top level categories were at a high level of abstraction, we decided to introduce a certain number of subcategories for each of them in addition to the simple top level category (named "neutral category"), reported in the section 4.1.1. There are two main reasons for this choice. The first is justifiable in terms of cognitive relevance. Categories more closed to the basic level are more natural and simple to describe. The second is related to the aim to investigate potential differences inside to the upper level categories in terms of attributes reported, identifying (in addition of attributes common to all different subcategories) also possible specific attributes for specific subcategories.

For each top level category we developed 6 (7 for the category EVENT) different scenarios one for

[8] The feature-listing task is a procedure for empirically deriving semantic feature norms, widely used to test theories of semantic representation that use semantic features as their representational currency.(for a detailed explanation of the method see [13]

Person	Organization	Event	Artifact	Location
politician	company	conference	product	tourist location
manager	association	meeting	artwork	city
professor	university	exhibition	building	shop
sports person	government	show	book	hotel
actor	agency	accident	article of clothing	restaurant
person	organization	event	object	location
		sports event		

Table 1
Subcategories used in the experiment.

each subcategory including the neutral category. By means of these scenarios we asked participants to produce a list of all attributes relevant for the specified category in order to obtain a unique profile of the entities that populate that category. There was no restriction in the number of attributes that could be reported. In table 4.1.2.1 we report the five lists of subcategories used in the experiment.

4.1.2.2. *Implementation* The experiment was conducted with a between-subjects design that is one subject was randomly assigned to only one combination of 5 scenarios (one subcategory for each top level category). This was required to eliminate interference between different scenarios. To guarantee a balanced distribution of subjects for each category we adopted a cyclic algorithm. Through the first cycle the algorithm selected randomly one scenario from each of the 5 lists and assigned the combination of scenarios to the first subject. In the second cycle the algorithm selected the scenarios immediately subsequent (in order) to those assigned in the previous step. When all items of one list were assigned, the algorithm began again from the completed list.

The experiment was conducted in three different versions: English (eng), Italian (it) and Chinese (chi) and was provided through the WWW. The subjects were invited (through email [9]) to participate in our online study. Once at this site, participants had to select the preferred language and were randomly assigned to an experimental condition, as described before; they then proceeded with 5 steps throughout the experiment: presentation, introduction, example, task and personal details.

Before starting the real task, participants were asked to read carefully the instructions which explained key terms used in the scenarios (for example the difference between "attributes" and "values" and the notion of "profile"). After that, a concrete example of the task was displayed. The domain of this example was deliberately chosen to be unrelated, to avoid that attributes reported as examples could interfere with the subsequent answers produced by subjects. For the real task, the five scenarios were presented in succession (the order was randomized between subjects). Finally, a personal detail page was presented. The aim was collecting information about provenance, age, gender, internet experience and semantic web experience of participants to use for further analysis. This part of the experiment was optional and could be skipped.

As incentive to participation we arranged a lottery to assign a prize [10] among the participants who completed the task. Subjects were free to decide whether to participate in the lottery or not. In case of participation, they were asked to submit their email address, but the anonymity of the experiment was guaranteed by making sure that this information was not aggregated with the experimental data [11].

4.2. *Results*

4.2.1. *Overview*

We collected data from 358 participants (159 for the English version, 194 for the Italian version and 5 for the Chinese version [12]), 181 of these were male, 102 female, 75 did not report gender information.

The average age of participants was about 31 years (considering only 285 subjects that actually provided age information). In table 12 we report the distribution of the number of subjects that specified their native country (262 out of 358), whereas in figure 1 we show the distribution in terms of Internet and Semantic Web experience, reported by 280 participants. From these self-evaluations it stands out that all subjects stated to have some knowledge in internet use and the majority of them reported "good" (117) or "expert"(134) knowledge. Differently, one-third of participants (102) reported none (54) or little knowledge (48) in the area of Semantic Web. Only 31 subjects defined themselves as

[9] To spread the participation request we submitted our post to mailing lists such as DBWorld or SIG-IRList

[10] We gave away a medium-priced MP3 player.

[11] Every participant was represented in our database by a numerical id, with the intent of tracing the combination of scenarios, the corresponding answers and the anonymous personal details. The email address was stored disconnected from these records.

[12] Because of the limited number of participants in the Chinese version in this paper we will present the results only of the Italian and English versions

Country	N	Country	N
Italy	141	United Kingdom	5
Brazil	19	Netherlands	3
Usa	14	Canada	3
Germany	14	Spain	3
India	11	Jordan	2
Pakistan	9	Malaysia	2
China	8	Mexico	2
Greece	6	Australia	2
Ireland	5	Switzerland	2
Others	21	N_{tot}	262

Table 2
Geographical provenance of participants.

experts in this area but a good part of participants reported "good" (85) or "average" (65) experience.

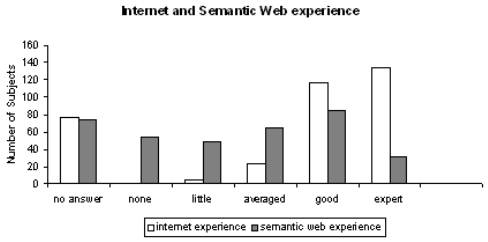

Fig. 1. Self-evaluation of the participants regarding Internet and Semantic Web Experience

Analyzing the data concerning the main task, participants reported on the average 5.16 attributes (median=5 and mode=5) [13]. From the values of mean, median and mode we can see that the distribution of our data is approximately symmetrical. For this reason we can reasonably assume that 5 is a good estimate of the number of attributes that subjects normally use to identify entities. Consequently we decided to present the results of the experiment, considering the first 5 attributes in our measures.

4.2.2. *Normalization*

As mentioned in the general description of the experiment, the peculiarity and the linguistic nature of the task made predictable a certain degree of variability in our data. To deal with this variability we normalized the data in three different steps: structural, morphological and semantic.

The first normalization step (structural) was performed mainly to report all answers in the form of lists of attributes. Indeed, although the instructions specified to insert one attribute per line in the specific form, some subjects disregarded this recommendation, using other break symbols (such as "," ";" "and" etc.) to separate the entries. Consequently, we had to implement a semi-automatic procedure to convert all the entries of our database in a standard form, splitting attributes so that the line number corresponded to the order of listing. This information will be extremely important for the future analysis on ranking. Moreover, in this first step, we checked the data to remove all typing errors.

The second normalization step (morphological) was finalized to report the attributes in a unique morphological form. For this purpose we removed articles, normalized the use of prepositions and the singular-plural inflections, we fixed the order for composed attributes (attributes which consists of two or more words).

Finally, the last normalization step (semantic), was conducted to aggregate attributes characterized by semantic overlaps (such as synonym expressions). In table 3 we report some examples of this preliminary phase. The number in brackets in the third column corresponds to the normalization step.

Attributes	Normalized form	Type of Normalization
name, address	name	splitting (1)
	address	
surename	surname	typing error (1)
the name	name	article erasing (2)
date of birth	birth date	order (2)
near cities	neighbouring cities	semantic overlap (3)
zip code	post code	semantic overlap (3)

Table 3
Normalization examples.

4.2.3. *Measures and Results*

The data analysis was conducted having in mind two different issues: the first deals with the intent to provide a small set of default entity types suggesting a possible (not fixed) description through attributes, the second pertains the possibility of exploiting the information enclosed in the description provided by

[13] If observations of a variable are ordered by value, the median value corresponds to the middle observation in that ordered list. The *median* value corresponds to a cumulative percentage of 50% (i.e., 50% of the values are below the median and 50% of the values are above the median). The position of the median is $\frac{n+1}{2}$, where n is the number of values in a set of data. The *mode* is the most frequently occurring value in a set of discrete data.

users to improve the efficacy of the entity matching algorithms.

To such issues correspond two different questions: firstly, which is the information most frequently specified by subjects when they provide descriptions of entity types investigated? secondly, which is the information more relevant to identify specific types of entities (distinguishing one type from others)? Trying to answer these questions we adopted two different measure: dominance and relevance (about the use of these measures in other contexts see [22,23,9]).

4.2.3.1. *Dominance* The problem of suggesting descriptions for types of entity at a high level of abstraction (corresponding to our top level categories) corresponds to identify a set of general attributes used by subjects across the subcategories of the same top level category. Aggregating the data from these subcategories we require a measure useful to evaluate the importance of an attribute f for the upper level category c.

To this purpose we adopted the dominance measure, that is a local measure that quantifies the importance of an attribute for a specific category. We can formalize the function of Dominance ϕ: $C \times F \rightarrow N$ in the following way:

$$dominance = \phi(c, f) = |\{s \in S : f \in F_s^c\}|$$

where S is the sample of subjects and F_s^c is the set of attributes listed by the subject s given the category c. In other words, the dominance ϕ of the attribute f for the category c is the cardinality [14] of the set constituted by all subjects that reported the attribute f for the category c. The dominance presents high scores when the attribute is frequently mentioned by subjects in defining the category.

In table 15 we report the dominance values (ϕ and $\phi\%$ [15]) for the two versions of the experiment. As previously remarked we fixed a threshold corresponding to 5 attributes, justified by the results on mean, median and mode. For the Italian version we report the original answers with the English translation in brackets. We note that the attribute more common across the categories is "name" which is the first attribute in two categories (Person and Organization) both in the Italian and in the English version and in the category Location but only in the English version. Moreover in the English version, "name" is present among the first 5 attributes in all categories. Personal attributes (name, surname, age, gender, birth-date) are most frequently reported to describe people. In addition to "name", organizations are identified in terms of spatial location (address, country) and type. Spatial (location) and time attributes (date, time) appear more relevant to describe events, whereas morphological and perceptual aspects (color, dimension, size, material) turn out to be more salient for the category Artifact. The most frequent attributes to describe locations are spatial (location, geographical coordinates, address, country).

4.2.3.2. *Relevance* Dominance does not provide information about the discriminatory power of an attribute f respect to a specific category c. If a user adopts a highly dominant attribute to describe an entity, we can not use this information to detect the presumptive category. The reason is that the dominance provide only a local evaluation of the importance of an attribute for the category without considering if the attribute is relevant also for others categories. Detecting those attribute which are dominant for a specific category but at the same time distinctive for it, is exactly the second aim of our research.

To identify attributes that correspond to this requirement we propose a measure, named relevance (k), that is the combination of two components: a local component (dominance) and a global component (distinctiveness). In the previous section we have formalized the first component. Now we pass to consider the second component.

The distinctiveness is a measure that quantify how much an attribute f is specific for a category c. When an attribute is used only in identifying one or few categories, its distinctiveness is high, whereas when it is used for many categories (or all) the distinctiveness score is low. The distinctiveness can be calculated as a function $\psi_d(f) : F \rightarrow [0, 1]$

$$distinctiveness = \psi_d(f) = 1 - \psi_s(f)$$

where $\psi_s(f)$ is a function of sharedness $\psi_s(f) : F \rightarrow [0, 1]$

$$sharedness = \psi_s(f) = \frac{|C[f]|}{|C|}$$

where $|C[f]|$ is the collection of the categories that have in common the attribute f and $|C|$ is the collection of all categories. If an attribute f is listed for all categories $\psi_d(f)$ is 0 and $\psi_s(f)$ is 1.

[14] number of the elements of a set

[15] $\phi\%=\frac{\phi}{N}$ where N=number of subjects

	English			Italian		
Category	**Attributes**	ϕ	ϕ%	**Attributes**	ϕ	ϕ%
Person	name	110	0.75	nome (name)	89	0.52
	age	49	0.33	età (age)	73	0.42
	gender	44	0.30	cognome (surname)	64	0.37
	birth-day	29	0.2	tipo (type)	56	0.32
	surname	24	0.16	data di nascita (birth-date)	34	0.19
		N= 145			N=171	
Organization	name	77	0.56	nome (name)	87	0.51
	location	37	0.27	tipo (type)	54	0.32
	country	34	0.24	luogo (location)	45	0.26
	address	31	0.22	scopo/i (aim/s)	44	0.26
	type	23	0.16	settore (sector)	23	0.13
		N= 137			N=168	
Event	location	116	0.79	luogo (location)	126	0.78
	date	69	0.47	data (date)	74	0.45
	time	64	0.43	tipo (type)	68	0.42
	name	49	0.33	ora (time)	57	0.35
	participants	40	0.27	partecipanti (participants)	39	0.24
		N= 146			N=161	
Artifact	color/s	46	0.32	colore/i (color/s)	74	0.44
	size	35	0.25	tipo (type)	60	0.35
	name	33	0.23	autore/i (author/s)	51	0.30
	title/s	29	0.20	dimensione/i (dimension/s)	36	0.21
	type	28	0.20	materiale (material)	35	0.20
		N= 140			N=168	
Location	name	86	0.59	luogo (location)	78	0.46
	country	50	0.34	nome (name)	73	0.43
	location	48	0.33	tipo (type)	57	0.33
	address	47	0.32	coordinate geografiche (geo co-ordinates)	35	0.20
	geo coordinates	43	0.29	numero abitanti (number of citizens)	29	0.17
		N= 145			N=169	

Table 4
Dominance for selected top-level categories.

The distinctiveness is a global measure because is transversal to all categories and in this sense it is category-independent and frequency-independent. This means that if we consider two different attributes f_1 and f_2, one used by all subject only in the category c_1 and the other used by only one subject only in the category c_2, their distinctiveness is identical ($\psi_d(f_1) = \psi_d(f_2) = 1/|C|$) regardless of the category and the number of subjects.

We can combine the two measures (dominance and distinctiveness) in a single measure, the relevance $k(c, f)$ [16], with the following formula:

$$k(c, f) = \phi(c, f) * \psi(f)$$

where $\psi(f)$ is a logarithmic transformation of the distinctiveness $\psi_d(f)$

$$\psi(f) = \ln \frac{|C|}{|C(f)|}$$

This measure can be adopted as an estimation of the contribution of an attribute f to identify a specific category c and, differently from distinctiveness, may be considered a concept-dependent measure. In other words, if the attribute is used by all (or the majority of) subjects to identify the category (high dominance) and is used only for that specific category (high distinctiveness), the relevance of the attribute for the category is consequently high. This means that the presence of that attribute is highly indicative (that is identifies with high probability) of the category considered. For example, the attribute "editor" is one of the most frequent attributes for the category *book* in both versions (it results in high values of dominance) and it is reported exclusively in the descriptions of that category (high values of distinctiveness). Combining dominance and distinctiveness, we obtain high values of relevance for this attribute when considered respect to the category *book*. Attributes with high values of relevance are highly informative for entity identification and entity matching. Continuing our example, consider the query q_1:<The Lord of the Rings and Allen & Unwin>. If we are able to recognize that "Allen & Unwin" is the name of an editor we ca use this information for the entity identification and matching, because the presence of the attribute "editor" suggests that the query refers most probably to the book rather than the movie that have the same title "The Lord of the Rings" (namely the same value for the attribute "title").

In tables A.1, A.2, A.3, A.4 and A.5 we report the measures of relevance, considering the first 5 attributes for each subcategories. In general we can notice that in every subcategory stand out some highly specific attributes that combine high-middle value of dominance coupled with high level of distinctiveness. Just to make some example, "party" for the subcategory politician, "faculties" for university, "sport specialty" for sport event, "editor" for book or "number of stars" for hotel. In addition to these specific attributes, every subcategory presents two or three of those attributes that we identified at the top of the lists of dominance. These attributes are less distinctive for the particular subcategory (that is they are widely shared by the subcategories inside their top level category but are not extensively shared by other subcategories resulting in intermediate values of distinctiveness) but compensate with very high values of dominance. For example, "surname" and "age" are attributes of this kind for the category Person. A case apart is represented by the attribute "name". As pointed in the section 4.2.3.1 this attribute is the most shared between the subcategories ($\psi_{it} = 0.93$, $\psi_{eng} = 1$). However if we consider carefully the nature of this attribute we can note that the presumptive meaning of it could be very different in different contexts. For the category Person, "name" can mean "first name" or a combination of "first name" and "surname" [17]. For the category Company, "name" can be synonym of "brand" and legal constraints regulate the organization name assignment at least in local contexts. Normally, for products "name" is associated to a class of objects (i.e. iPhone 3G) with the same features and not to a single object (my iPhone). In the light of these differences, we decided to consider the attribute "name" distinct for the five top level categories. Using this expedient, we found that the attribute "name" appear nearly in all subcategories among the 5 most relevant attributes. In support of our methodological choice of aggregating data across the subcategories to obtain a list of general attributes as suggestions for entity description,

[16] We point out the similarity of the relevance measure with the tf-idf measure often used as term weighting approach in information retrieval and text mining [21].

[17] We suppose that the tendency of considering "name" as the combination of "first name" and "surname" is more likely for English speakers. Indeed in the Italian version of the experiment 63 participants (out of 89 that reported the attribute "name") listed "name" and "surname" as two different attributes, whereas in the English version only 24 subjects (out of 110) listed the two attributes combined.

we found that the most relevant attributes for the neutral categories correspond well enough to those found by means of the dominance measures obtained from aggregated data sets. The reason that why we adopted the aggregation strategy is primarily due to the size of the sample (the neutral category samples have about one sixth of subjects in comparison to the aggregated samples).

4.3. *Applications*

As briefly sketched in the introduction, the driving factors for this research were two-fold: entity representation, and entity matching.

4.3.1. *Entity Representation*

We can directly apply our findings to the way we represent entities in the Entity Name System. This will have special influence for future evolutions of the system, where are going to devise ways to foster a certain convergence between how users *describe* entities, and how they *search* for entities.

Some of the client applications that are using the ENS today have been updated to give the user a selection of our top-level types listed in Sect. 4.2, to manually classify an entity to be created. Subsequently, we provide the properties found to be most important for this entity type as a proposed "default schema" to the user, that can be manually filled with values.

As a second step, the knowledge we gained from investigating the co-occurence of attributes enables us to work on a way to remove the manual classification step in favour of automatic classification. This is a more complex scenario that requires knowledge-based methods, which in our ENS use cases we are going to deal with. Imagine the keywords "Costa Forza Italia": for a human (with some background knowledge), it is relatively easy to understand that we are describing a person called "Costa" who is member of the political party "Forza Italia", and not – what would be another imaginable interpretation – a stretch of coast in Italy that is named "Forza". The following steps facilitate such an automatic classification process:

(i) Through the use of Named Entity Recognition (NER) functionality, which will be able to detect that "Forza Italia" is a political party; thus, we can tokenize the description into two parts: $t_1 = Costa$, which is still unknown at this point, and $t_2 = Forza\ Italia$, which we have just classified.
(ii) Relying on our findings of this paper, we can assume that "political party" is an attribute only relevant for politicians.
(iii) With the use of a background ontology (or a simpler structure that formalizes the results presented here), we can know that politicians are of type PERSON.
(iv) Based on our findings, we know that the most relevant attribute of PERSON is "name", so we can argue that the token t_1 is probably the name of the entity.

As a result, we can (a) provide a schema proposal to describe the entity, and (b) pre-populate the schema with the values already provided. We expect this to have significant positive influence on the "cleanliness" of data, and on the convergence between entity representation and entity matching, as we will explain in the following.

4.3.2. *Entity Matching*

The second application area that we are directly interested in is entity matching, i.e. the attempt to return the single one entity that a user was (most probably) looking for when searching the ENS.

There are two ways how the research findings presented here can be applied to this problem: in a straight-forward manner, to take into serious account which descriptive attributes are more relevant for distinguishing entities, and a "backward" manner, by making inferences about the desired type of entity from a given search term.

The first case can be exploited by giving higher weights to the more relevant attribute types when ranking search results. To give a brief sketch, for example, as we have illustrated in Sect. 4.2, the "name" attribute usually has a high relevance; so for a search term x and two entities $E_1 = \{name = x\}$ and $E_1 = \{place\_of\_birth = x\}$, it can be argued that E_1 is the better match, because the search term appears in the more relevant feature.

A second way to make use of our findings is related to the issue of developing an advanced matching algorithm for a problem, by guessing the *type* of entity that is to be matched, based on co-occurance of descriptive attributes. We are attempting to mimic human behaviour of "understanding" what is the intention behind a bag of search words, by applying the following steps:

(i) First, we can perform automatic classification

based on co-occurence of attributes, similarly as explained before. The only difference is that now we are classifying a query string, to infer what kind of entity a user is searching for.

(ii) With the help of a thesaurus-based approach, we can approximate the "name" field in an entity description in different natural languages or representations ("nombre", "nome", "http://xmlns.com/foaf/0.1/name", ...).

(iii) Finally, we can give assign an appropriately higher weight to this field when matching entities, as described before.

In the light of the result that "name" seems to be by far the most relevant attribute to describe entities, we do expect matching requests for entities to also reflect this phenomenon. We thus plan to directly apply the findings presented here to work on algorithms that work on co-occurence of attributes similar to the example described above. Such algorithms will concentrate on (a) classifying what type of entity a matching request is most probably aiming at, and (b) relating search tokens to the most probable attributes of this entity type (i.e. which of the tokens most probably is the name of an entity, and which on is just "description"). To the best of our knowledge, this represents a novel approach, and we expect this to help us achieve higher-precision results without the a-priory knowledge (or enforcement) of any specific representational schema for entities. A first step in this direction is presented in the next section, where we describe an approach for entity matching that bases on parts of the facts discussed here.

5. A Novel Approach for Entity Matching

For underlining the usefulness of the results and their proposed applications described in the previous sections, we have developed a novel, proof-of-concept approach for entity matching in a large entity repository, called *Name-feature Matching*, or *nfm* for short. In the following, we will describe the underlying Name-feature Score, a first algorithmic implementation, an experimental setting in which we tested the approach, and first test results.

5.1. *The Name-feature Score*

To present a ranked list of results that match a query (or to decide about an "optimal" match), we require a score that serves as parameter for ranking, which expresses the closeness of an individual entity to the query, relative to all other candidates.

In our setting, both the representation of a query (Q) and entity (E) is in the form of *features*, which consist of name/value pairs that are independent in content and size (i.e. they don't necessarily share a vocabulary or schema). Formally speaking, Q and E are sets of *features*; a feature is a name/value pair $< n, v >$.

v_i^Q is a notation we introduce to denote the *value* part of the i-th element of Q. Likewise, v_j^E is the value part of the j-th element of E. Similarly, n_i^Q denotes the *name* part of the i-th element of Q, and n_j^E denotes the name part of the j-th element of E.

First, we define $fsim(Q_i, E_j)$, a function that computes the similarity of two features Q_i, E_j, taking into account the similarity of the value parts, as well as the cases where the name parts denote *names* in the linguistic or ontological sense; $fsim$ is defined as follows:

$$fsim(Q_i, E_j) = sim\left(v_i^Q, v_j^E\right) * \begin{cases} 2 * v * n, & \text{for name}(n_i^Q), \text{name}(n_j^E), \text{id}\left(v_i^Q, v_j^E\right); \\ 2 * n, & \text{for name}(n_i^Q), \text{name}(n_j^E); \\ * \quad v * n, & \text{for } n_i^Q = \emptyset, \text{name}(n_j^E), \text{id}\left(v_i^Q, v_j^E\right); \\ n, & \text{for } n_i^Q = \emptyset, \text{name}(n_j^E); \\ 1, & \text{otherwise.} \end{cases} \quad (1)$$

The computation of $fsim$ relies on the following functions and parameters:

$sim(x, y)$: a suitable similarity measure between x and y [18];

$name(x)$: a boolean function indicating whether the feature x denotes one of the possible "names" of the entity (as discussed in Sect. 4.3.2, and further described in 5.2);

$id(x, y)$: the identity function, true if the value parts of x, y are identical;

n : the factor to which a name feature is considered more important than a non-name feature;

v : the extra weight attributed to the occurance of value identity, i.e. $id(x, y)$.

To compute the Name-feature Score, which finally expresses to which extend E is similar to Q, we proceed as follows.

[18] Possible candidates for implementation are Levenshtein Distance [8] as used for our experiments, or any other algorithm that produces a suitable similarity measure between strings.

Let maxv(V) be a function that computes the maximum value in a vector [19].

We then span the matrix M of feature similarities between Q and E, defined as

$$M := (fsim\,(Q_i, E_j))_{|Q|\times|E|} \rightarrow \mathbb{Q} \geq 0$$

with $fsim$ as defined above, and $|Q|, |E|$ being the number of elements of the vectors Q and E, respectively.

The Name-feature Score nfs is defined as the sum of all the maxmimum similar feature combinations between Q and E:

$$nfs\,(Q, E) = \sum_{i=1}^{|Q|} maxv(M_i) \qquad (2)$$

Example: let us imagine a query with two features and an entity described by three features ($|Q| = 2$ and $|E| = 3$); then the computation of nfs first spans the matrix M using the $fsim$ function for each feature combination(with example values as below), computes the vector V of the maxima, and sums up all elements of V into a single score:

$$\begin{pmatrix} 0.3 & 1.8 & 3.0 \\ 0.1 & 4.2 & 0.5 \end{pmatrix} \mapsto \begin{pmatrix} 3.0 \\ 4.2 \end{pmatrix} \mapsto 7.2$$

There are two main characteristics of this approach: (1) if a feature is detected to denote a name, it receives extra weight. A possible implementation of a method to detect this circumstance is given in the next section; (2) the formula spans a matrix of the features of Q and E, and takes into account only the maximum similar feature combinations; this prevents an effect of "spamming", in which an entity with a large number of features could accumulate a higher score by sheer number of similar features, compared to another entity that would in reality be a better match, but has a lower number of features.

Even though in the description of the Name-feature Score we have implied to compare a query against entities, due to the compatible representations that we assume, it can be applied without modification in scenarios of de-duplication or entity linkage which compares entities with entities.

5.2. *The NameFeatureMatch Implementation*

The *nfm* algorithm accepts as input the representation of the entity that is supposed to be found (i.e. the "query"), consisting of a set of name/value pairs $N = \{< n, v >\}$. It produces as output an ordered list $R = \{< id_1, nfs_1 >, \ldots, < id_n, nfs_n >\}$ of entity identifiers id, together with their individual score nfs, ordered by nfs.

Structurally, the algorithm is rather straightforward: it iterates over the features of the query for every element of a set of candidate entities $C = \{E\}$, building a matrix M as described before, selecting the maximum of each row, building the final score for the entity, and sorting the resulting list of Name-feature Scores.

The main feature of the algorithm however, is to detect whether a feature is in fact a *name* or not, in order to be able to decide which of the cases in Eq. 1 has to be applied. To this end, we have implemented a class called NameFeatureDetector for the $name(x)$ function in Eq. 1, which employs a combination of regular expressions together with a small, multilingual [20], hand-crafted thesaurus, which lists the most common natural language and formal vocabulary elements that are used to denote *names*. Based on this easily extensible and very efficient thesaurus, we produce a set of regular expressions that we run against the features, and return a boolean answer that is suitable for the calculation of nfs.

As for the rest of the ingredients necessary to calculate nfm, we employed the Levenshtein string similarity measure [8] for $sim(x, y)$, implemented $id(x)$ in a straight-forward manner, and set $n = v = 2$.

5.3. *Experimental Dataset*

To evaluate our approach, we ran a set of example queries against the current population of the Okkam ENS, which consists of approximately 593,000 entities, the demographics of which are reported in Table 5.

The entities are stored in the Okkam Ens [2], which holds one *Entity Profile* for each entity [19]. In the Ens, search for an entity is accelerated by pre-selection through index retrieval, which delivers a maximum number of candidates per query, currently with a ceiling of 1,000 entities. The matching algorithm is subsequently run on these candidates in-memory.

To establish a benchmark for the performance of the approach, we extracted 200 random queries from

[19] Trivially defined as $maxv\,(V) = max_{i=1}^{n}\,(V_i)$, with n being the size of V.

[20] Currently covering English, Italian and German.

Data Source	Number of Entities
Locations (subset of Geonames)	77,541
Semantic Web Conference Entities	1,017
Organizations (from Wikipedia)	99,325
People (from Wikipedia)	415,105
Total	**592,988**

Table 5
Demographics of ENS Population, November 2008

the Ens query log. These queries represent a mixture of those entered manually by a human (e.g. through our web-search interface [21]), and ones created automatically by Ens-enabled applications, such as the ontology editor Protégé for which we have introduced a plugin extension [3]. The search queries were syntactically de-duplicated [22] , and queries that represented illegal syntax were removed. Finally, for every query, a human expert manually established the "correct" Ens identifier of the *intention* behind the query, to the extent that this was possible. Queries whose intention was unclear even to the expert were eliminated from the set.

5.4. *Evaluation*

For this first evaluation of the prototypical implementation, we only concentrated on the top-1 success rate of our algorithm for cases where the respective query actually has a definite match in the Ens, which left a number of 67 queries to work with. This means that only the highest-ranked result is considered and compared with the standard, whereas the occurence of the correct result at any other than the top position is counted as failure [23] .

To give a comparative evaluation, we have run the same experiment also on three other approaches: first, as a base-line we use directly the results that the index retrieval of the Ens provides (hence labelled "STORE"); second, a naive algorithm that relies strictly on string matching, as implemented for an early prototype of the Ens [24] (labelled "NAIVE"); and third, an adapted implementation of the Group Linkage algorithm [18] (labelled "GROUP"). This latter one constitutes the current approach that is in use in the ENS; we thus have to assume that this approach has been selected for the good results it produces in the given setting. A more comprehensive evaluation of our approach against existing ones is discussed in the Future Work section.

	NFM	GROUP	STORE	NAIVE
Avg. Runtime (ms)	2969	3613	657	3642
Queries processed	67	67	67	67
Top-1 Successes	31	23	16	4
Top-1 Misses	36	44	51	63
Top-1 Success Rate	46%	34%	24%	6%
Improvement over baseline	94%	44%	0%	-75%

Table 6
Performance comparison between nfm and other relevant approaches.

The results, reported in Table 6, show an improvement achieved by nfm compared to the other approaches. The improvement with respect to the Group Linkage implementation is especially significant because it also shows that nfm does not introduce a runtime penalty [24] .

6. Future Work

The relevance measure proposed in our analysis does not consider the ranked nature of participant responses. However, we believe that the *order* in which attributes are listed can convey information about the attribute relevance. Participants of our experiment were requested to list all attributes that they considered relevant to identify each category. We assume that subjects have followed some order

[21] `http://www.okkam.org/ens/Search.jsp`
[22] I.e. "London" and "london" are considered different queries, which helps us to investigate in the robustness of our approach.
[23] This may appear rather strict, but is an important aspect when dealing with systems that rely on high-precision batch processing, such as the annotation of whole datasets with Ens identifiers. In such cases, the batch processor has no other choice than to accept the top element of the result set, and the outcome of the process stands or falls based on the top-1 performance of the matching algorithm employed to produce the results.
[24] Please note that the absolute number of milliseconds presented here are a mere indicator, to give a rough overview and comparison. All four experiments were run concurrently on the same machine(a XEN Linux virtual machine, on a Dual-Core AMD Opteron(tm) Processor at 2,6GHz, with 4GB of RAM), which is not the actual target hardware. On production hardware with a natively running operating system, the Ens reaches response times of around 600ms including data transfer over the Internet.

when they listed the attributes, starting with attributes more salient which occurred first to them and so on until attributes less important. Models that ignore this crucial feature presume that later mentioned attributes represent the category equally well as earlier ones, providing a partial representation of the cognitive salience of attributes for the category. To model this further aspect of the semantic relevance we are developing a new measure that considers the order of attributes given, in addition to the other two components (dominance and distinctiveness) of the model worked with in this article.

In order to test which of these models best fits the real data we have to perform some control experiments. For example we are planning to perform a reversed experiment in which participants are asked to identify the type of entity corresponding to a description, in which the combination of attributes is manipulated on the basis of relevance measures. To test the effectiveness of the model in predicting the accuracy of the response, we have to test whether descriptions with higher global relevance (that is the mathematical combination of the relevance values of the individual attributes) are more likely to suggest the correspondent entity type than descriptions with lower relevance.

On the methodological side, one of the most important contributions of our research consists of importing the experimental paradigm of the feature-listing task in a new research context. This paradigm, largely used in cognitive science and neuropsychology to study categorization and category-specific semantic impairment, has been adopted to investigate how subjects represent types of entities by means of attributes. The nature of the task requires a remarkable abstraction effort from the subjects. Usually people are not used to think in terms of *types* of attributes when they describe entities in their Web search activities, but instead they tend to specify values of attributes.

Consequently we have to be sure that attributes collected by means of this paradigm correspond to those really used by humans. In order to test the ecological validity of our results it is important to investigate the real use of attributes in other more "natural" contexts, such as the formulation of queries in actual Web search, or the choice of attributes in a more "descriptive" context like the Wikipedia infobox [25].

Another interesting aspect that we tried to investigate has been suggested by the results of cultural psychology studies that found cultural differences in semantic intuitions about reference [10] and categorical judgments [16]. In the light of this evidence we tried to investigate the cultural difference in using identification attributes, especially to highlight possible differences between Eastern and Western users. Unfortunately the data from the Chinese version are insufficient to allow us to make some inferences. A further experiment will collect more data from the Chinese or other Eastern Country and provide us significant data for the comparison.

Finally, we are planning to juxtapose the work described in this paper, as well as the experiments described as further work so far, with the "reality" of the Ens. This will be effected by means of analyzing query logs as well as entity descriptions. Log analysis will provide us with valuable insight about how people search for entities. Reviewing entity descriptions on the other hand are obviously highly suitable to perform a comparative study with the findings presented here. This will all be performed after the Ens has been in operation for some time, because of the fact that its initial population will be largely created automatically from existing structured data sources, which do not have the collaborative character we are looking for in this context. Both approaches offer the advantage of being able to work on a relatively large amount of data that can hardly be gathered otherwise. Additionally, because queries as well as entity data in the Ens are stored together with temporal information, we can direct further research towards the temporal *evolution* of entity descriptions.

As for the implementation of entity matching approaches, the short-term plans are to improve the nfm algorithm. As evident from the first evaluation presented here, negative cases, i.e. whether a query that has no results in the Ens is correctly answered with a zero result, were not considered. This is left for future work, because it relies on an evolution of the Name-feature Score: currently, nfs is an open-ended score that is built by accumulating points, which is sufficient for ranking in the context we described, because it makes results of one run of the algorithm comparable against each other; however, to be able to detect true or false negatives, we need to normalize the score in order to make results

[25] An infobox on Wikipedia is a consistently-formatted table which is present in articles with a common subject to provide summary information consistently between articles or improve navigation to closely related articles in that subject.

comparable *across different runs*, so that in the end by experimenting we can define a threshold under which the algorithm can assume that none of the results is good enough to be considered a match, and return an adequate response. To improve the quality of the algorithm itself, we are planning on the one hand to improve the thesaurus approach for detecting the name feature, and on the other hand to run suitable optimization techniques which will help to find more optimal settings for the two parameters n and v. Finally, an important next step is to become less self-referential in the evaluation of the performance of the algorithm, and compare it with more relevant approaches. In the mid-term, we plan to work on an additional approach that involves type detection based on features, and that gives weights not only to the name features, but also to others, based on their significance for identifying an entity.

7. Conclusion

We expect that an agent that interacts with the Entity Name System (for example searching for an entity) has first to perform a cognitive operation that consists in building a mental representation of a specific entity, and selecting from this representation a subset of features that he or she evaluates relevant enough to identify that entity. This means that the attributes once specified in the corresponding values provide a description that has to be used to verify the presence of the entity in the repository (entity matching), distinguishing it from the others. Unfortunately, what seems obvious for a human being can be obscure for an automatic system. To facilitate the matching processing, it is necessary to implement some heuristics that exploit supplementary information, such as the kind of entity searched or the attributes most likely used to describe it. This kind of information is often not in explicit form. In this paper we described an empirical investigation about entity descriptions, provided by a large sample of subjects performing a feature-listing task. We proposed a measure of relevance to analyze the results, and have applied the findings to the specific problem of entity matching in a large entity repository. With the evaluations we performed, we were able to show that an approach that takes into account the cognitive point of view of entity representation by humans can provide an improvement over other relevant approaches.

Acknowledgements

This work is partially supported by the by the FP7 EU Large-scale Integrating Project **OKKAM – Enabling the Web of Entities** (contract no. ICT-215032). For details, visit `http://www.okkam.org`.

The authors would like thank the UNITN Okkam team for their support: Sven Buschbeck for implementing the web-based poll and helping analyze the collected data; Mariana Zerega, Liu Xin and Stefano Bortoli for helping with creating the benchmark; Daniel Giacomuzzi for implementing the testing framework.

Appendix A. Relevance Tables for Entity Types

In this section we report the tables containing the top-5 features for each category (Person, Event, etc.), including their individual subcategories which have been used in the experiment described in Sect. 4. As mentioned before, the data presented here are an excerpt from the comprehensive study available in [1].

PERSON				
Category	**Attributes (eng)**	k	**Attributes (it)**	k
Politician	party	65.78	partito (party)	35.63
	name	31.20	orientamento politico (political view)	28.67
	gender	16.42	nome (name)	22.99
	position/s	14.60	età (age)	21.02
	age	11.30	cognome (surname)	18.06
Manager	name	26.27	azienda di appartenenza (company)	21.93
	experience/s	11.68	nome (name)	21.34
	reports	10.30	cognome (surname)	18.06
	department/s	9.12	esperienze (experiences)	9.12
	occupation	7.01	titolo di studio (education)	9.85
Professor	university	37.77	materia di insegnamento (teaching matter)	54.94
	name	34.48	nome (name)	21.34
	publications	19.19	istituzione in cui insegna (institution where teaches)	20.60
	research area	17.17	pubblicazioni (publications)	17.17
	department	14.60	cognome (surname)	13.14
Sportsperson	type of sport	49.34	specialità sportiva (sport specialty)	54.82
	name	31.20	nome (name)	22.99
	team	17.17	età (age)	17.31
	birth date	11.50	cognome (surname)	11.50
	gender	14.78	data di nascita (birth date)	9.85
Actor	name	26.27	films	34.34
	films	19.19	nome (name)	24.63
	birth date	11.50	esperienze (experiences)	25.54
	gender	9.85	età (age)	19.79
	awards	6.14	cognome (surname)	16.42
Person	name	31.20	nome (name)	32.84
	gender	22.99	cognome (surname)	27.92
	birth date	18.06	luogo di nascita (birth place)	18.25
	occupation	14.01	professione (occupation)	18.68
	religion	13.74	data di nascita (birth date)	16.42

Table A.1
Relevance for PERSON

ORGANIZATION				
Category	**Attributes (eng)**	k	**Attributes (it)**	k
Company	name	24.63	nome (name)	36.12
	ceo name	8.22	numero di dipendenti (number of employees)	14.78
	business type	8.19	fatturato (turnover)	12.29
	profits	7.01	capitale sociale (share capital)	10.96
	revenue	6.87	produzione (output)	10.30
Association	name	21.34	nome (name)	27.91
	objective/s	16.42	associati (members)	27.47
	members	11.68	scopo/i (objective/s)	11.61
	activity	11.68	numero iscritti (number of members registered)	7.01
	date of foundation	9.12	funzioni (functions)	6.87
University	name	26.27	facoltà (faculties)	30.91
	number of students	19.19	nome (name)	22.99
	faculty/ies	16.35	numero di studenti (number of students)	17.17
	courses	13.70	corsi (courses)	16.35
	department/s	9.12	docenti (professors)	13.74
Government	name	11.49	orientamento politico (political view)	16.38
	head	9.34	durata (duration)	13.74
	members	9.34	partito/i (party/s)	10.96
	party	8.22	ministeri (ministries)	10.30
	leaders	8.22	ministri (ministers)	10.30
Agency	name	21.34	nome (name)	26.27
	number of employees	7.44	numero di dipendenti (number of employees)	8.21
	president	6.87	clientela (clients)	6.87
	specialization	4.67	settore (sector)	4.53
	profit/s	4.67	scopo/i (objective/s)	4.36
Organization	name	21.34	nome (name)	27.91
	business type	6.14	scopo/i (objective/s)	12.34
	objective/s	4.93	membri (members)	11.68
	character/s	4.67	settore (sector)	9.05
	head	4.67	data di fondazione (date of foundation)	8.21

Table A.2
Relevance for ORGANIZATION

EVENT				
Category	**Attributes (eng)**	k	**Attributes (it)**	k
Conference	name	22.32	argomento (topic)	25.54
	organizers	12.77	relatori (speakers)	16.45
	date	12.37	partecipanti (participants)	16.42
	chair/s	10.30	data (date)	11.13
	sessions	10.30	necessità (needs)	10.30
Meeting	time	29.76	argomento (topic)	20.07
	date	19.79	ora (time)	17.86
	topic/s	18.43	partecipanti (participants)	16.42
	participants	16.08	data (date)	11.13
	agenda	13.74	luogo (location)	7.01
Exhibition	name	11.90	argomento (topic)	21.89
	time	8.93	durata (duration)	11.50
	date	7.42	artisti partecipanti (artists)	10.96
	start date	7.01	espositori (exhibitors)	10.30
	end date	4.93	titolo (title)	9.85
Show	actors	17.17	data (date)	19.79
	name	13.39	attori (actors)	18.68
	producer/s	10.30	durata (duration)	18.06
	time	7.44	ora (time)	16.37
	director/s	7.01	titolo (title)	14.78
Accident	time	20.83	persone coinvolte (people involved)	25.69
	date	11.13	entità coinvolte (entities involved)	24.04
	people involved	10.96	danni (damages)	13.74
	participants	7.42	dinamica (dynamics)	13.74
	causes	7.01	veicoli coinvolti (vehicles involved)	10.30
Event	time	17.86	data (date)	18.55
	date	17.31	partecipanti (participants)	13.14
	name	10.41	nome (name)	12.77
	participants	9.89	durata (duration)	8.21
	repetition	6.87	ora (time)	7.44
Sports event	type of sport	30.15	specialità sportiva (type of sport)	27.41
	stadium	10.30	nome (name)	14.59
	date	12.37	data (date)	13.60
	time	7.44	luogo (location)	7.01
	winners	6.87	ora (time)	5.95

Table A.3
Relevance for EVENT

ARTIFACT				
Category	**Attributes (eng)**	k	**Attributes (it)**	k
Product	manufacturer	21.02	utilizzo (use)	22.52
	price/s	14.84	nome (name)	18.06
	name	14.78	prezzo (price)	14.24
	use	14.33	colore (color)	11.13
	warranty	10.30	marca (brand)	9.34
Artwork	artist/s	30.91	autore (author)	37.77
	creation date	18.43	luogo (location)	24.57
	style	14.01	stile (style)	18.68
	material	10.95	data di creazione (creation date)	17.17
	author	7.01	tecnica (technique)	17.17
Building	architect	20.60	numero di piani (number of floors)	37.77
	number of floors	17.17	luogo (location)	32.76
	height	14.90	metratura (mq)	20.60
	name	13.13	altezza (height)	16.37
	architectural style	10.30	anno di costruzione (date of creation)	13.74
Book	author/s	46.71	editore (editor)	61.81
	publisher	44.64	numero di pagine (number of pages)	54.94
	ISBN	35.63	autore (author)	44.34
	year of publication	27.47	titolo (title)	36.13
	number of pages	24.04	anno di pubblicazione (publication date)	30.91
Article of clothing	gender intended for	24.04	taglie (sizes)	72.11
	color	17.08	marca (brand)	35.03
	material	16.42	colore (color)	33.39
	style	16.35	tessuto (fabric)	27.47
	fabric	13.74	modello (model)	17.17
Object	shape	16.42	colore (color)	24.74
	color	13.29	materiale (material)	22.99
	id	11.68	forma (shape)	22.99
	value	10.30	funzione (function)	21.02
	name	9.85	peso	18.25

Table A.4
Relevance for ARTIFACT

LOCATION				
Category	**Attributes (eng)**	k	**Attributes (it)**	k
Tourist Location	name	18.06	attrazioni (amenities)	21.93
	attractions	13.74	nome (name)	11.49
	geo coordinates	11.50	possibilità di svago (amusements)	10.96
	price/s	4.95	numero di abitanti (number of inhabitants)	9.34
	area	4.67	posizione geografica (geographical position)	9.34
City	population	25.69	numero abitanti (number of citizens)	56.05
	name	21.34	coordinate geografiche (geo coordinates)	23.72
	geo coordinates	11.50	nome (name)	19.70
	number of people	9.34	regione (region)	18.25
	language/s	9.12	clima (climate)	14.01
Shop	products sold	13.74	merce trattata (type of products sold)	48.08
	name	13.13	nome (name)	31.20
	quality	5.48	orario (time)	30.15
	owner/s	5.42	numero di dipendenti (number of employees)	11.50
	price/s	4.95	luogo (location)	8.77
Hotel	name	32.84	numero di stanze (number of rooms)	27.41
	number of rooms	13.74	numero di stelle (number of stars)	27.41
	rating	10.96	nome (name)	24.63
	amenities	10.30	servizi (services)	12.45
	number of stars	10.30	categoria di appartenenza (category)	10.84
Restaurant	type of cuisine	52.08	nome (name)	21.34
	name	34.48	orario (time)	19.19
	chef	17.17	specialità (specialty)	13.74
	kind of food	13.74	piatti tipici (typical food)	10.30
	price/s	13.60	prezzo/i (price/s)	9.49
Place	geo coordinates	42.70	coordinate geografiche (geo coordinates)	29.19
	name	21.34	indirizzo (address)	16.82
	continent	13.74	nome (name)	11.49
	elevation	10.30	altitudine (altitude)	10.96
	distance from see	8.22	continente (continent)	8.22

Table A.5
Relevance for LOCATION

References

[1] Barbara Bazzanella, Heiko Stoermer, and Paolo Bouquet. Top Level Categories and Attributes for Entity Representation. Technical Report 1, University of Trento, Scienze della Cognizione e della Formazione, September 2008. http://eprints.biblio.unitn.it/archive/00001467/.

[2] Paolo Bouquet, Heiko Stoermer, Claudia Niederee, and Antonio Mana. Entity Name System: The Backbone of an Open and Scalable Web of Data. In *Proceedings of the IEEE International Conference on Semantic Computing, ICSC 2008*, number CSS-ICSC 2008-4-28-25, pages 554–561. IEEE Computer Society, August 2008.

[3] Paolo Bouquet, Heiko Stoermer, and Liu Xin. Okkam4P - A Protégé Plugin for Supporting the Re-use of Globally Unique Identifiers for Individuals in OWL/RDF Knowledge Bases. In *Proceedings of the Fourth Italian Semantic Web Workshop (SWAP2007), Bari, Italy, Dec.18-20, 2007*, December 2007. http://CEUR-WS.org/Vol-314/41.pdf.

[4] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16, 2007.

[5] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007.

[6] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with dolce. In *Lecture Notes In Computer Science, Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, volume 2473, 2002.

[7] Hector Garcia-Molina. Pair-wise entity resolution: overview and challenges. In Philip S. Yu, Vassilis J. Tsotras, Edward A. Fox, and Bing Liu, editors, *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, Arlington, Virginia, USA, November 6-11, 2006*, page 1. ACM, 2006.

[8] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.

[9] L. Lombardi and G. Sartori. Models of relevant cue integration in name retrieval. *Journal of Memory and Language*, 57:101–125, 2007.

[10] E. Machery, R. Mallon, S. Nichols, and S.P. Stich. Semantics, cross-cultural style. *Cognition*, 92:B1–B12, 2004.

[11] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and Al. Oltramari. *WonderWeb Deliverable D18 Ontology Library (final)*. 2003.

[12] C. Matuszek, J. Cabral, M. Witbrock, and J. DeOliveira. An introduction to the syntax and content of cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering.*, Stanford, March 2006.

[13] K. McRae, G.S. Cree, M.S. Seidenberg, and C. McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instrument and Computers*, 37:547–559, 2005.

[14] G.A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97, 1956.

[15] I. Niles and A. Pease. Towards a standard upper ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, Ogunquit, Maine, October 17-19, 2001.

[16] R.E. Nisbett. Culture and systems of thought: holistic versus analytic cognition. *Phychological Review*, 108:291–310, 2001.

[17] A. Oltramari, A. Gangemi, N. Guarino, and C. Masolo. Restructuring wordnet's top-level: The ontoclean approach. In *n K. Simov (ed.) Workshop Proceedings of OntoLex'2, Ontologies and Lexical Knowledge Bases*, Las Palmas, Spain, May 27, 2002.

[18] Byung-Won On, Nick Koudas, Dongwon Lee, and Divesh Srivastava. Group linkage. In *ICDE*, 2007.

[19] Themis Palpanas, Junaid Chaudhry, Periklis Andritsos, and Yannis Velegrakis. Entity data management in okkam. In *SWAE, Turin, Italy*, September 2008.

[20] E. Rosch, C.B. Mervis, W. Gray, D.Johnson, and Boyes-P. Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[21] G. Salton and Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–123, 1988.

[22] G. Sartori and L. Lombardi. Semantic relevance and semantic disorders. *Journal of Cognitive Neuroscience*, 16:439–452, 2004.

[23] G. Sartori, G. Negri, I. Mariani, and S. Prioni. Relevance of semantic features and category specificity. *Cortex*, 40:191–193, 2004.

[24] Heiko Stoermer. *OKKAM: Enabling Entity-centric Information Integration in the Semantic Web*. PhD thesis, University of Trento, January 2008. http://eprints.biblio.unitn.it/archive/00001394/.