

The Impact of Semantic Handshakes

Lutz Maicher

University of Leipzig, Augustusplatz 10-11, 04109 Leipzig, Germany
maicher@informatik.uni-leipzig.de

One of the key challenges for the breaking through of the semantic web or web 2.0 is global semantic integration: if two proxies in different models represent the same thing in the “real world” they should become mergeable. The common top-down approach to semantic integration is the enforcement of centralised ontologies or PSI repositories. This top-down approach bases on a overly optimistic premise: the success of one universal vocabulary enforced by a central authority. This paper proposes a bottom-up approach. A semantic handshake is the decision that two terms from different vocabularies can be used to identify the same “thing”. If these local decisions are broadcasted, global integration can be achieved without any ontological imperialism. Within this paper this hypothesis is investigated by simulations. We show that if the majority of proxies describes its identity only by *two* different public known terms, global integration is almost achievable at the large scale.

1. The Challenge of Semantic Integration

One central challenge in each kind of modeling is this subtle identity relationship between a “thing” in the real world and it’s proxies in the models. This relationship has to be made explicit if models from different sources should become mergeable. If the identity relationship is made explicit, it could be decided whether two proxies from different models are representatives of identical “things” in the real world.

The general and widely adopted approach for creating mergeable models is the definition, evangelization and usage of ontologies, schemas or PSI repositories. Such standardized vocabularies can be used to express the identity relationship between a proxy and the thing it represents in the real world. If two different proxies in different models should represent the same thing, the model creators can use the identical term provided by the central ontology to express the identity relationship. In the case all model creators use the same ontology, global integration is achievable. Global integration means, that all proxies in diverse models representing the same thing become mergeable.

We assume that this ontology approach bases on an overly optimistic premise: the success of a top-down approach, the definition and enforcement of an universal vocabulary by a centralized authority. We expect that in practice global integration is not achievable by trying evangelizing one centralized vocabulary.

In this paper we will discuss a bottom-up approach to come closer to the goal of global integration. The premise of our approach is that for expressing the same

2 Lutz Maicher

relationship a lot of different “terms” are defined in diverse ontologies or vocabularies. In practice these different terms are used simultaneously. Instead of evangelizing one universal term out of the universe of terms, our approach bases on the usage of local integration decisions. A local integration decision is the commitment of the model creator that term A and term B from different vocabularies can be used to express the same identity relationship. If this local integration decision is broadcasted, all proxies originally only using term A become mergeable with all proxies originally only using term B. This paper investigates the impact of these distributed local semantic agreements which we will call *semantic handshakes*.

In Topic Maps the advent of exchange protocols (like TMRAP [Ga05] or TMIP [Ba05]) allows the request and exchange of proxies having the same identity from distributed, heterogeneous models. A peer requests from a remote peer whether proxies with the same identity are available. In that case, the remote peer responds with the appropriate proxy and the requesting peer can merge (parts of) the received model in its local model. If the network is requested the next time, new terms for expressing the identity can learned from the requested peers can be used to improve the request.

We will show by simulations, that if the majority of proxies describes its identity by *two* different public known terms (i.e. two different results from swoogle [1]), the existing terminological diversity can be preserved and global integration is almost achievable. We assume that the bottom-up approach of semantic handshakes does even better fit the requirements of the practice as the enforcement of centralised vocabularies.

The remainder of this paper is the following. In section 2 the theoretical background for the simulations is given. It bases on the identity approach introduced in Topic Maps [TMDM, TMRM], the international industry standard for information integration. In the section 3 the simulation design is described in full detail. In section 4 different experiment series based on the simulation design are described and discussed. In section 5 related research is described and section 6 summarizes the findings of the experiment series.

2. Theoretical Background

Topic Maps are a modeling method which enforces the disclosure of the identity relationship of each proxy [DNB06]. This means if a proxy should represent “Bernd Hilfreich” (facts assigned to this proxy are statements about the person which is called Bernd Hilfreich) the proxy has to disclose its identity by at least one string.¹ According to the theory two proxies have to be merged, if they have the same identity. (In Topic Maps terms they represent the equal subject.)

Decisions about subject equality are straightforward: if two proxies have at least one pair of strings representing the identity in common, both are considered to have the same identity and both have to be merged. Merging of two proxies is well defined

¹ The following description of the identity mechanism in Topic Maps is a mixture and a simplification of the TMRM [DNB06] and one of its legends, the TMDM [TMDM].

and leads to one proxy having the union set of all properties of the original proxies. The following example illustrates the identity approach²:

```
[id = "id1"; identity identifiers = {"I1", "I2"}; names = {"Bernd Hilfreich"}]
[id = "id2"; identity identifiers = {"I2", "I4"}; names = {"Bernd"}]
[id = "id3"; identity identifiers = {"I5"}; names = {"Meyers, Jim"}]
```

According to the rules defined above, the first two entities are considered to have the same identity and both have to be merged. The third entity is considered to have a different identity and rest untouched:

```
[id = "id1,id2"; identity identifiers = {"I1", "I2","I4"};
      names = {"Bernd Hilfreich", "Bernd"}]
[id = "id3"; identity identifiers = {"I5"}; names = {"Meyers, Jim"}]
```

For the simulations a slightly different identity and merging mechanism will be used: a proxy does not have any other properties than one proxy identifier (to refer to the proxy as object of the model) and a set of *comparable* identity identifiers for disclosing the identity of the proxy. Subject equality of two proxies holds, if the intersection of their sets of identity identifiers is not the empty set. In that case, the set of identity identifiers of *both* proxies will become the union of their sets of identity identifiers. In contrast to the integration model above, all proxies continue to exist and only the sets if their identity identifiers will be merged and will grow in time. Global integration is achieved, if all a proxies representing the same subject have the identical set of identity identifiers.

To illustrate the impact of the local semantic handshakes, the example given above should be viewed from a distributed perspective. All three proxies *id1*, *id2* and *id3* should be considered to be part of different distributed models. All of these proxies request all known remote models, whether proxies with the same identity are available. As result, the set of identity identifiers of *id1* and *id2* become merged. Thus the local decision that the identity of *id1* can be described by "I1" and "I2", and the independent local decision that the identity of *id2* can be described by "I2" and "I4" will be broadcasted then. The next time *id2* will request remote models, the request can be improved by "I1". In the next sections, the enormous impact of this simple effect towards a bottom-up standardization through distributed, local semantic handshakes is investigated with simulations.

3. Simulation Design

This section describes the simulation design in detail. The simulation setting is completely implemented in Java and well documented. Both, implementation and documentation are available at [2] and can be used for further experiment series. The remainder of this section is organised as follows. The first parts define some terminological specifications. In the subsequent parts the process implemented in the simulation setting is described in more detail.

² The property „names“ is only introduced for illustration purposes.

Experiment Series, Experiment, Test, and Merge Roundtrip

Each simulation is an experiment series, which consist of a sequence of parameterised experiments. Each experiment is a sequence of tests. Each test is a sequence of merge roundtrips. In this document these terms are used according the following intensions:

Experiment Series. An experiment series is a sequence of parameterised experiments. Usually, one parameter iterates (in example the number of different identity identifiers which are “known” in the world) in a given range.

Experiment. An experiment is a sequence of tests. Because the setup of a test environment is a stochastic process, the results of experiments are means of measures observed in a sequence of tests.

Test. A test is one process as described below. According to the given parameters, all proxies are created and identity identifiers assigned. Within a test a specified number of merge roundtrips is executed.

Merge roundtrip. A merge roundtrip is the following process: for each proxy in E it is decided whether there are other proxies available in E which have to be merged with the given proxy.

Terminological Specifications

We will define E as a set of proxies e_i which have by definition the same identity. For example E might be the set of all available proxies of the type “person” or E might be the set of all available proxies of the individual “Bernd Hilfreich”. Each proxy e_i has a unique proxy identifier which is used to refer to this proxy³. Additionally, each entity e_i discloses its identity by a non empty set I_i of identity identifiers. Identity identifiers are comparable and it is always decidable whether two identity identifiers are equal or not. The set T_i of a proxy e_i consists of the proxy identifiers of all proxies which are considered to have the same identity as e_i (identity equality has already hold).

Two proxies e_i and e_j will be considered as equal (identity equality holds) if

$$(1) \quad e_i = e_j \Leftrightarrow I_i \cap I_j \neq \emptyset$$

If proxy e_i is equal to proxy e_j merging will create two proxies e_i' and e_j' in E' with the following characteristics:

$$(2) \quad I_i' = I_j' = I_i \cup I_j$$

$$(3) \quad T_i' = T_j' = T_i \cup T_j$$

³ For clarity, the value of the index i will be the value of the proxy identifier. In example, e_{id1} is the proxy with the proxy identifier idl . The same holds for all variables, like I_i and T_i .

The premise of the simulation design is that all proxies in E have the same identity. But this can only be globally exploited by information systems, if identity equality is detected between all entities in E . In terms of the simulation design, global integration is achieved if T_i of all entities e_i in E is equal to E^4 :

$$(4) \quad \forall e_i \in E \mid \text{card}(T_i) = \text{card}(E)$$

After these terminological specifications, in the following the process implemented in the simulation setting is described.

Initialisation of a Test

In the first step of a test, E has to be initialised. The variable $\text{card}E$ defines the number of proxies which have to be created.⁵ To each proxy a unique proxy identifier is assigned. The variable $\text{distributionNbrOfII}$ defines the distribution of the *numbers* of identity identifiers which will be assigned to the proxies. (In the section “Defining Distributions” of [2] the definition of a distribution is described in detail). According to this variable, for each e_i the number of identity identifiers which have to be assigned to is calculated stochastically.

Afterwards, a value for each identity identifier has to be created. This will be done stochastically according to the distribution defined by the variable distributionII . The variable nbrOfDifferentII the number of different identity identifiers which are known in the world. Therefore, the *value* of an identity identifier is a number in $[1, \text{nbrOfDifferentII}]$.

Example. The distribution for the values of the identity identifiers might be defined as follows $\{[0.8, 1.0], 6\}$. This is equivalent to the lottery that with a probability of 80% an identity identifier gets the value 1, 2 or 3. In the same time, with a probability of 20% an identity identifier gets the value 4, 5 or 6. This means, that half of the six possible identity identifiers are widely used and the other half of the six possible identity identifiers is rarely used.

Executing a Merge Roundtrip

A test is a sequence of merge roundtrips (the number of merge roundtrips is defined by the variable $\text{nbrOfMergeRoundtrips}^6$). In a merge roundtrip for each proxy e_i in E identity equality to all other entities in E is decided according to (1). If identity

⁴ This holds iff e_i is contained in T_i (otherwise T_i should consist of $\text{card}(E)-1$ proxies). The comparison of the set cardinality is allowed because T_i only consists of elements from E .

⁵ Experiments have shown, that $\text{card}E$ partially influences the result. If $\text{card}E$ is less than a threshold both $\text{card}(T)$ and $\text{clouds}(E)$ changes simultaneously with $\text{card}E$. In the case $\text{card}E$ exceeds this threshold both values are not influenced by its changes. In all cases, the threshold is less than $\text{card}E=100$. Therefore, in all following experiments $\text{card}E$ is set to 100.

⁶ Through the connectedness of all proxies, the results does not change after the second merge roundtrip. If the connectedness of proxies would become a stochastic process, too, more merge roundtrips become necessary.

equality holds e_i' and e_j' will be created in E' according to (2) and (3). After the merge roundtrip all e_i in E which have counterpart in E' will be replaced by this e_i' .⁷

Analysing an Experiment Series

To get statistically valid measures, each experiment is a sequence of tests with the same instantiation parameters. This is necessary due to the stochastic nature of the initialisation process. The number of tests in an experiment is defined by *nbrOfTests*.

For comparing the influence of parameters within an experiment series different measures have to be calculated. These measures specify the size and nature of the *integration clouds* which emerge in the tests. An integration cloud is a set of proxies within E where identity equality is considered. Global integration is achieved, if there exist only one integration cloud. This cloud has the size *cardE*.

card(T). This measure depicts the average size of an integration cloud in E after a test. Formally, it is the weighted average cardinality of T_i of all e_i in E . The algorithm is implemented in `Simulation.getAverageCardT()` [2].

Note. This measure favours large integration clouds because the size s of a cloud is the weight for the weighted average. Given three integration clouds (one of size 98, and two of size 1) *card(T)* is 96,06.

clouds(E). This measure depicts the number of different integration clouds in E . Formally, it's the maximal number of T_i in E which have empty intersections. The algorithm is implemented in `Simulation.getNbrOfClouds()` [2].

To evaluate an experiment, the mean of all tests' *card(T)* and the mean of all tests' *number of cluster(E)* are the appropriate measure. Within an experiment series, these measures for parameterised experiments are compared.

4. Results of the Experiment Series

This section introduces and discusses different experiment series. Starting from a scenario where a global ontology is enforced, different parameters influencing the global integration are investigated. Besides the implementation and the documentation of the simulation setting, [2] provides the protocols of all experiment series. We urgently recommend the consultation of this additional material.

Global ontology

If the overly optimistic premise holds and global ontologies or global PSI repositories are enforceable, further experiment series might be not necessary. In that case, I_i of all e_i will consist of only one element: the globally unique identity identifier. After *one* merge roundtrip *card(T_i)* of each entity e_i is *cardE* and *clouds(E)*

⁷ The separation of E and E' is necessary to avoid further mergings within one merge roundtrip. For example, if a proxy gets a new identity identifier through merging, new merging opportunities might occur. Through separating E and E' these new opportunities will be executed in the next merge roundtrip.

is one. Global integration is reached. But the premise of our research is that this top-down approach is overly optimistic.

A completely heterogeneous world without any semantic handshakes

The counterpart of the enforcement of global ontologies or global PSI registries is a completely heterogeneous world. In that case, each e_i gets its own *globally unique* identity identifier and no semantic handshakes are done. Obviously, the global integration defined by (4) can never be achieved. After each merge roundtrip, $card(T)$ will be always 1, and $clouds(E)$ will be always $cardE$.

A partly heterogeneous world without semantic handshakes

In a first step, the constraint of *globally unique* identity identifiers for each proxy will be softened. In the following experiment series, to each proxy e_i only one identity identifier will be assigned. But, instead of being globally unique, the identity identifier assigned to each e_i is a value randomly chosen (according to a uniform distribution $distributionII=\{1.0\}$) in the range $[1, nbrOfDifferentII]$. (From a given set of identity identifiers one identity identifier for each proxy is drawn.) As a result, two different proxies will get the same identity identifiers with a certain probability (depending on $nbrOfDifferentII$). In the experiment series $exp01$ ⁸ shown in Figure 1 $nbrOfDifferentII$ iterates from 5 to 100.

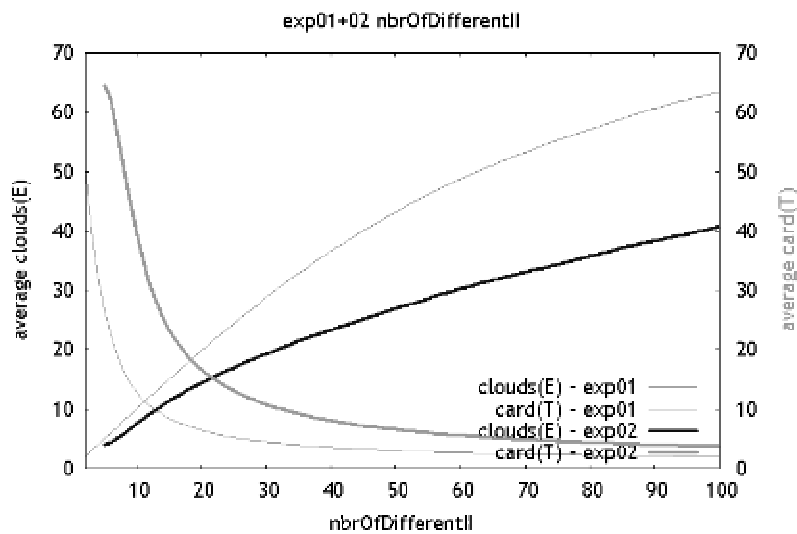


Figure 1 exp01+02 Iterating $nbrOfDifferentII$ in $[5,100]$
general parameters: $cardE=100$, $distributionNbrOfII=\{1.0\}$
specific parameter exp01: $distributionII=\{1.0\}$
specific parameter exp02: $distributionII=\{0.8,0.9,0.95,1.0\}$

⁸ The detailed protocol of experiment series $exp01$ is available at [3].

In the experiment series *exp02*⁹ shown in Figure 1 the parameter *distributionII* is set to {0.8,0.9,0.95,1.0}. This means, that the identity identifiers are not drawn according to a uniform distribution. Instead of, some identity identifiers are more popular than others.

The results of the experiment series *exp01* show, that for small *maxII* the number of resulting integration clouds *clouds(E)* is equal to *nbrOfDifferentII*. If five different identity identifiers are available in the world, five separate clouds of nearly identical size will appear.

The more *nbrOfDifferentII* increases, the more the average number of *clouds(E)* is less than *nbrOfDifferentII*. This has a simple rationale: if for 100 proxies an identity identifier has to be chosen, this is similar to a hundredfold repetition of drawing an identity identifier from the given set of identity identifiers. If the cardinality of this set is 5, *clouds(E)* is only less than five in the case, if after 100 trials one of the five given identity identifiers is not drawn one time. This is not expectable. But if the cardinality of the set of identity identifiers is i.e. 80, there is a significant probability that one of these 80 identity identifiers is not drawn in 100 trials.

The experiment series *exp02* shows the influence of the distribution of the identity identifiers. In this series the identity identifiers are drawn according to a distribution with some popular and a lot of unpopular identity identifiers. The results improve significantly. The size of the resulting clouds increases due to the fact that popular identity identifiers imply bigger clouds. But even the number of clouds *clouds(E)* decrease significantly due to the strengthening of the effect discussed related to experiment series *exp01*. Nevertheless, great fragmentation rests in *E*. (This is problematic because we assume that experiment series *exp02* reflects the current state in the practice: there are a lot of different terms, some of them are more popular and all of them are used simultaneously.)

The impact of semantic handshakes in a partly heterogeneous world

In the following the impact of semantic handshakes will be investigated in detail. A semantic handshake is done, when two different identity identifiers are assigned to one proxy. In that case, the distribution of the *number* of identity identifiers which will be assigned to proxy have to be changed. Changing *distributionNbrOfII* to [{0.3,1.0},2] means, that 30% of all proxies will get one identity identifier randomly drawn from the universe of identity identifiers and 50% of all proxies will get two randomly drawn identity identifiers. Starting the iteration of *distributionNbrOfII*={a,1.0} with *a*=0.0 means, that all proxies will get two different randomly chosen identity identifiers. In contrast *a*=1.0 means, that to all proxies only one randomly drawn identity identifier will be assigned. (This situation is equal to experiment series *exp01*.)

Experiment series *exp03*¹⁰ shown in Figure 2 bases on the assumption that the assigned identity identifiers are uniformly distributed (*distributionII*={1.0}). If all proxies get two different identity identifiers (*distributionNbrOfII*={0.0,1.0}), the results are very impressive: average *clouds(E)*=4 and average *card(T)*=92. Due to the

⁹ The detailed protocol of experiment series *exp02* is available at [4].

¹⁰ The detailed protocol of experiment series *exp03* is available at [5].

semantic handshakes, more than 92% of all proxies are accumulated within one cluster. Around a maximum of 3 further semantic handshakes are sufficient to achieve global integration.

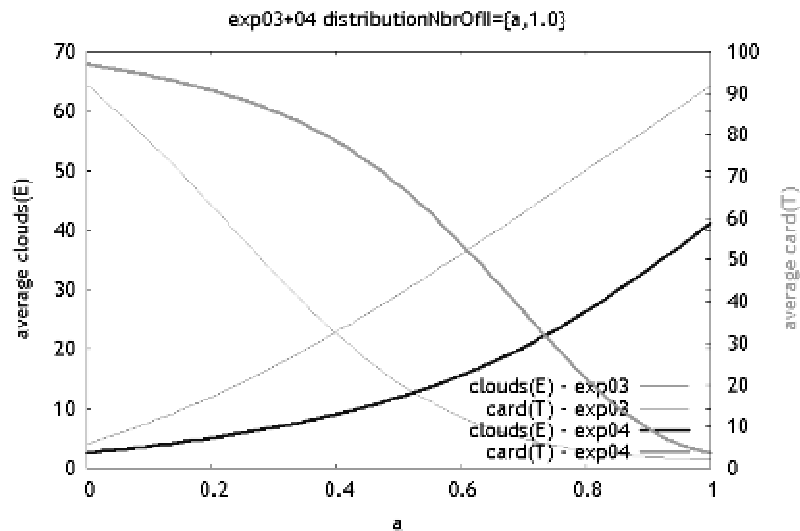


Figure 2 exp03+04 Iterating a in $distributionNbrOfII=\{a,1.0\}$ in $[0.0,1.0]$
 general parameters: $cardE=100, nbrOfDifferentII=100$
 specific parameter exp03: $distributionII=\{1.0\}$
 specific parameter exp04: $distributionII=\{0.8,0.9,0.97,1.0\}$

Furthermore, in the experiment series $exp04$ ¹¹ shown in Figure 2 the existence of popular identity identifiers is assumed. The value of the variable $distributionII$ is changed from $\{1.0\}$ to $\{0.8,0.9,0.97,1.0\}$. In that case, both values increase significantly: $clouds(E)=2.5$ and $card(T)=97.0$. In fact, more than 97% of all proxies are integrated within one integration cloud. Only around 1.3 further semantic handshakes are sufficient to gain global integration.

It is remarkable, that these results are similar to reducing the number of possible identity identifiers ($nbrOfDifferentII$) to a very small number (according to the findings of $exp01$ and $exp02$). Whereby reducing the number of possible identity identifiers have to be enforced by a centralised authorisation, the concept of semantic handshakes is based on decentralised, autonomous decisions.

We assume that only a part of all proxies will barrow a semantic handshake. Therefore, the results for $distributionNbrOfII=\{0.0,1.0\}$ should be a interpreted as a best world scenario. To be more realistic, a view to the development of the result quality during the iteration is necessary.

From this perspective, Figure 2 does reveal the influence of popular identity identifiers in experiment series $exp04$. In the case, where only to the half of all proxies a semantic handshake is assigned ($distributionNbrOfII=\{0.5,1.0\}$), the results are still impressive: $clouds(E)=10.7$ and $card(T)=75.9$. This means, that there still exist an

¹¹ The detailed protocol of experiment series $exp04$ is available at [6].

integration cloud which consists of more than 75% of all proxies. In contrast, in experiment series *exp03* with uniformly distributed identity identifiers the results are less convenient: $clouds(E)=14.0$ and $card(T_i)=28.8$.

The influence of the diversity of identity identifiers

When investigating the impact of semantic handshakes in the experiment series *exp03* und *exp04*, the diversity of the available identity identifiers was big ($nbrOfDifferentII=100$). As already shown in the experiment series *exp01* and *exp02* (by iterating over $nbrOfDifferentII$), a lower diversity of the available identity identifiers has a significant impact to the quality of the results. In the following the connection of semantic handshakes and the diversity of the available identity identifiers should be investigated.

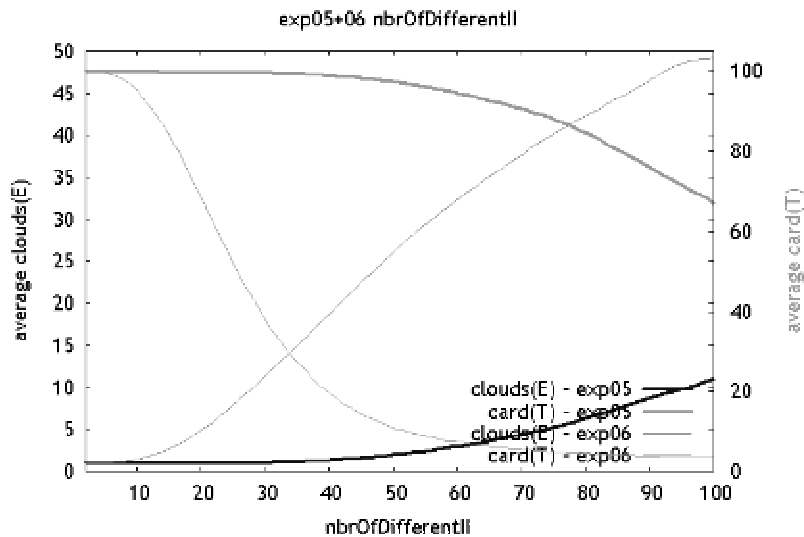


Figure 3 exp05+06 Iterating $nbrOfDifferentII$ in [2,100]
general parameters: $cardE=100, distributionII=\{1.0\}$
specific parameter exp05: $distributionNbrOfII=\{0.2,1.0\}$
specific parameter exp06: $distributionNbrOfII=\{0.8,1.0\}$

In the experiment series *exp05*¹² shown in Figure 3 semantic handshakes are assigned to the majority of proxies: $distributionNbrOfII=\{0.2,1.0\}$. The results are very impressive: even if 40 different identity identifiers exist, global integration will be achieved. It has to be outlined, that the top-down approach using centralised ontologies tries to achieve this global integration by evangelising one universal identity identifier. These findings illustrate the impact of semantic handshakes very well.

¹² The detailed protocol of experiment series *exp05* is available at [7].

But even if semantic handshakes are only assigned to a minority of proxies, the quality of the results increase significantly. In the experiment series *exp06*¹³ the variable *distributionNbrOfII* is set to {0.8,1.0}. In the case only 40 different identity identifiers exist, the results are: *clouds(E)*=18.9 and *card(T)*=16.0. This is a dramatic decline in contrast to *exp05*. But in contrast it is a significant improvement in contrast to *exp01*, where (ceteris paribus) no semantic handshakes are assigned: *clouds(E)*=36,9 and *card(T)*=3.5.

More semantic handshakes versus more sophisticated semantic handshakes

Within this section it will be investigated, how the result quality can be further improved. Two different strategies are evaluated. The first strategy is assigning more than one semantic handshake to a proxy. In fact a proxy gets more than two identity identifiers. The second strategy is to exploit the popularity of identity identifiers. The question is whether two very popular identity identifiers or one popular and one non well-known identity identifiers should be assigned to a proxy.

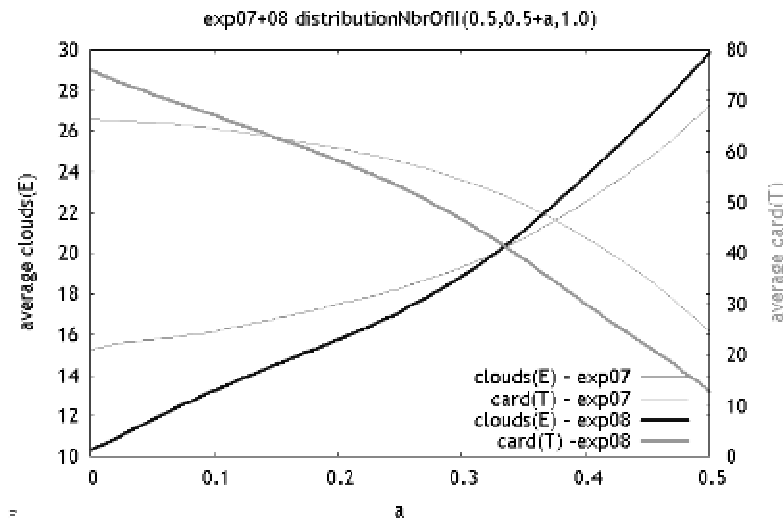


Figure 4 exp07+08 Iterating a in $distributionNbrOfII=\{0.5,0.5+a,1.0\}$ in $[0.0,0.5]$

general parameters: $cardE=100$, $nbrOfDifferentII=100$,

specific parameter exp07: $distributionII=\{1,0\}$, $distributionII2=\{0.8,0.9,0.97,1,0\}$

specific parameter exp08: $distributionII=\{1,0\}$, $distributionII2=\{1,0\}$

Both experiment series *exp07*¹⁴ and *exp08*¹⁵ shown in Figure 4 illustrate the impact of a second semantic handshake (the first improvement strategy). The starting point of these series is a big diversity of available identity identifiers ($nbrOfDifferentII=100$). To all proxies one randomly drawn (according to a uniform distribution) identity

¹³ The detailed protocol of experiment series *exp06* is available at [8].

¹⁴ The detailed protocol of experiment series *exp07* is available at [9].

¹⁵ The detailed protocol of experiment series *exp08* is available at [10].

identifier is assigned. Furthermore it is assumed, that only to the half of the proxies (at least one) semantic handshake is assigned. If $a=0.5$ ($distributionNbrOfII=\{0.5,1.0,1.0\}$) to all of these proxies only one semantic handshake is assigned. If $a=0.0$ () to all of these proxies two semantic handshakes are assigned. The results auf *exp07* and *exp08* show, that assigning more than one semantic handshake enhances the quality of the results significantly.

The second improvement strategy was the qualified choice of identity identifiers. In *exp08* all identity identifiers are drawn according $distributionII=\{1.0\}$. The experiment series *exp07* follows a different schema: the first and the third (if assigned) identity identifiers are drawn according $distributionII=\{1.0\}$, but the second identity identifier (if drawn) is drawn according $distributionII=\{0.8,0.9,0.97,1.0\}$. In fact, the second identity identifier is always a popular one.

Figure 4 shows that the strategy of *exp07* yields better results if the minority of proxies get a third identity identifier. Otherwise, the strategy of *exp08* might be better. These results are due to the fact, that half of all proxies do only have one identity identifier and this identifier is randomly chosen from the universe of identity identifiers. These equivocal findings imply the following strategy: identity identifiers should be assigned like the majority did it in the past. If popular identity identifiers can be exposed, these popular identity identifiers should be used. Otherwise, a random choice might be appropriate.

5. Related Research

The problem of scaling shared vocabularies is part of the research field called emergent semantics [ACC⁺04]. From the perspective of our research, a relevant work in the context of emergent semantics is [ACH03]. While Aberer et al. focus on the problem of achieving the decision about the semantic handshakes, this paper evaluates the premise of approaches like emergent semantics: the suitability of bottom-up approaches.

The idea of semantic handshakes is influenced by Gladwells “The tipping point” [G100]. He revealed that local interactions can have significant global impact if a certain threshold is exceeded. This holds for semantic handshakes, too. If a majority of proxies does disclose local semantic handshakes, global integration can be achieved without centralised authorization.

The web 2.0. bases on distributed tagging using folksonomies. These folksonomies explicitly do not rely on central authorizations. However, using semantic handshakes these tags become mergeable at the large scale which allows a diversity of new applications using these tag data.

Semantic handshakes are means for terminological standardization and vocabulary evolution in a bottom-up fashion. The development of vocabularies in self-organizing systems is investigated by Steels [St96].

6. Discussion

The experiment series have shown, that the semantic handshake approach might be very appropriate to achieve the goals discussed in the introduction: preserving the existing terminological diversity and achieving global integration.

To achieve these goals the following guidelines for proxy creators can be derived from the findings:

- add always at least *two* different identity identifiers to one proxy (disclosure of the semantic handshake) and
- use popular identity identifiers.

Both design rules are that much important that we propose to make them to a central part of model engineering methods, i.e. a topic maps engineering. Spreading the idea of semantic handshakes around by evangelising this modelling techniques, a majority of proxies will disclose semantic handshakes and their impact becomes significant. Furthermore, using popular identity identifiers leads de facto to terminological standardisation. The results of *exp05* can be interpreted as follows: if a proxy uses one of the 40 most popular identity identifiers (*cardE=100*), it will be definitely part of the main integration cloud.

Otherwise it is obviously, that observed or created “specialities” (very seldom identity identifiers) should be made public by assigning it to a proxy which has already some public identity identifiers.

Naturally, the simulation design does represent a “best case” scenario in which all proxies are connected and exchange their identity identifiers immediately. Due to this connectedness stable integration clouds are always established after two iterations (merge roundtrips). This speed does not reflect the real world, but we assume that if the simulation achieves global integration, real life applications will come close to it in finite time. We propose further experiment series where the connectedness of proxies become a stochastic and in time changing property.

Nevertheless, the proxies (or their creators) must be able to exchange information about their identity by any means. For example, making all public topic maps querable by TMRAP or TMIP, an enormous pool of identity identifiers occur and can be exploited for the purposes of semantic handshakes. We assume that this approach might be more practicable than defining and maintaining centralised PSI repositories. The same holds for RDF or other metadata repositories.

Naturally, if the handshakes are incorrect the approach of broadcasting semantic handshakes implies some problems. In this case, wrong but enormous integration decisions can be inferred (this is of great impact especially if two big complementary clouds of entities are touched). The main problem is, that semantic handshakes once submitted into *E*, might lead to a chain of new integrations. It will be complicated to trace back to the situation before the incorrect semantic handshake was done. To avoid incorrect semantic handshakes only identity information from trusted sources should be used. In that case, fraud will harder be broadcasted. This strategy does not avoid taking over incorrect semantic handshakes which were made accidentally by a trusted source.

Summarised, we assume that knowing and using the impact of semantic handshakes semantic integration on the large scale is achievable. It seems to be a

more realistic way than the attempt of the evangelisation of one universal vocabulary by a central authority.

7. References

- [ACC⁺04] Aberer, K.; Catarci, T.; Cudré-Mauroux, P. ; et al.: *Emergent Semantic Systems*.
- [ACH03] Aberer, K.; Cudré-Mauroux, P.; Hauswirth, M.: *The chatty web: Emergent Semantics Through Gossiping*. Proceedings of the 12th International World Wide Web Conference, Budapest (2003).
- [Ba05] Barta, R.: *TMIP, A RESTful Topic Maps Interaction Protocol*. In: Proceedings of Extreme Markup Languages 2005, Montreal, (2005).
- [DNB06] Durusau, P.; Newcomb, S. R.; Barta, R.: *Topic Maps Topic Maps Reference Model, 13250-5*.
- [Ga06] Garshol, L. M.: *TMRAP – Topic Maps Remote Access Protocol*. In: Proceedings of First International Workshop on Topic Maps Research and Applications (TMRA'05), Leipzig; Springer LNAI 3873, (2006).
- [GI00] Gladwell, M.: *Der tipping point. Wie kleine Dinge Grosses bewirken können*. Berlin-Verlag, Berlin (2000).
- [St96] Steels, L.: *Self-organising vocabularies*. In: Proceedings of Artificial Life V, (1996).
- [TMDM] ISO/IEC: Topic Maps – Part 2: Data Model. Latest version available at: <http://www.isotopicmaps.org/sam>
- [1] <http://swoogle.umbc.edu/>
- [2] <http://www.informatik.uni-leipzig.de/~maicher/sh/sh.htm>
- [3] <http://www.informatik.uni-leipzig.de/~maicher/sh/Protokolle/exp01.htm>
- [4] <http://www.informatik.uni-leipzig.de/~maicher/sh/Protokolle/exp02.htm>
- [5] <http://www.informatik.uni-leipzig.de/~maicher/sh/Protokolle/exp03.htm>
- [6] <http://www.informatik.uni-leipzig.de/~maicher/sh/Protokolle/exp04.htm>
- [7] <http://www.informatik.uni-leipzig.de/~maicher/sh/Protokolle/exp05.htm>
- [8] <http://www.informatik.uni-leipzig.de/~maicher/sh/Protokolle/exp06.htm>
- [9] <http://www.informatik.uni-leipzig.de/~maicher/sh/Protokolle/exp07.htm>
- [10] <http://www.informatik.uni-leipzig.de/~maicher/sh/Protokolle/exp08.htm>