A Trust Model to Estimate the Quality of Annotations using the Web

Davide Ceolin
VU University
Amsterdam
dceolin@few.vu.nl

Willem Robert van Hage
VU University
Amsterdam
wrvhage@few.vu.nl

Wan Fokkink
VU University
Amsterdam
wanf@few.vu.nl

ABSTRACT

The objective of this paper is to propose a method for estimating the trust level of annotations of professional media. We develop a model based on subjective logic and semantic web technology, and subsequently test this on a sample set of annotations from a natural-history museum.

Keywords

Trust, Annotations, Subjective Logic, Semantic Web

1. INTRODUCTION

1.1 Context: professional media

Professional media include all digital media, such as images, video and audio used for professional purpose. The term "professional" refers to all media used for business (both profit and non-profit) such as in entertainment, culture, science and product catalogs. In professional media there is a strong *supplier* point of view. Typical suppliers are TV broadcasters, musea and digital libraries. In this paper we focus on examples of musea as professional media providers.

Authority and quality are two key issues for professional media. Musea invest heavily in building expertise on the items in their collection. Metadata creation is therefore a core and knowledge-intensive activity in the management of professional media. A museum plays the role of authority within its field of competence. This means that it has the responsibility to keep and protect the artifacts it owns (including digital representations of the work), to guarantee their preservation over time.

However, musea often have large collections which are only partly properly catalogued. They simply do not have the resources to cover the complete collection. For this reason musea are looking with great interest at the current Web 2.0 trend of "tagging". An example is the Steve Museum [2]. But tagging of museum artifacts is a different ball game when compared to tagging of say family photos. Tagging of museum artifacts is a task that requires a high level of skill for the annotator, because of the precision and quality needed for tags. In this context, quality and precision of the tags is what assures the authority to keep its authoritative position. Both the selective amount of skills needed to annotate properly and the consequences of a low quality tagging are important issues for professional media.

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.
.

1.2 The concept of trust

The concept of trust is well-known in sociology and law. Trusting is a risky activity which involves at least three entities: trustor, trustee, and beneficiary. In this activity, the trustor delegates some tasks and discloses some possibly sensitive information to the trustee. The trustor is the subject delegating or sharing information (that is, trusting), in which a certain degree of risk is involved. The aim of the task is to allow the beneficiary, that is the end user, to take advantage of the result of the delegation or disclosure process. To do this, the trustor needs to rely on the trustee, because of the particular skills or information owned by this subject. The process of sharing and delegating involves a certain degree of risk, since it implies the loss of direct control by the trustor.

Camp [5] defines trust as the overlap of the following three facets:

Security: The act of disclosure and sharing of sensitive information directly implies that enough guarantees should be offered regarding the security level of these data. Security is a broad term; in this context it refers mainly to the intentional damage that the trustor may suffer by taking part in the process of trusting. These damages may be inflicted by either the trustee or a third party being able to somehow interfere in the process.

Privacy: While security focuses on intentional damage, privacy centers on unwanted disclosure of sensitive information beyond the boundaries of the trust process.

Reliability: Beyond the sharing needs, delegation plays a key role within trust. Delegation of tasks is needed when the trustor is not able, for many possible reasons, to deliver the task. These reasons may include the particular skills needed to deal with the tasks, or the workload implied, for instance. A reasonable belief in the trustee's reliability is therefore essential to allow the trustor to trust.

This characterization emphasizes possible risks involved in trust, but the last point encloses also the key motivation for trust, that is, the possible, hoped gain for the trustor. In the next section we will situate trust within the context of professional media.

1.3 Trust and professional media

The role of the trustor is played by the authority, *i.e.* the museum or the TV broadcaster. This authority owns the media and wants to share it with the public, which is the beneficiary, without running the risk of compromising its authoritative position. Situated between trustor and beneficiary is the trustee: the actor that allows the delivery of content, for instance by properly annotating it.

As the process can take place in different physical environments, security and privacy issues may mainly derive from interaction with the external world. For instance, when the process implies that the content is delivered through a partially protected channel, then this may imply threats from a security point of view. Moreover, since business relations are involved, also privacy is fundamentally important. These aspects depend mainly on the infrastructure used and on the kind of relations established by the different parties (*e.g.* commercial contract), and may therefore imply *ad hoc* solutions. We focus on the reliability aspect.

Reliability evaluation is anyhow necessary because the trustor's authoritative position may be seriously damaged by wrong annotations, since annotations are provided to the users together with the content, within its authoritative space (*e.g.* the museum website). Trustworthy metadata provision is a key activity on which *e.g.* musea run their business. Their public desires to trust them, and this trustworthiness is achieved through the delivery of trustworthy information, in particular trustworthy metadata. Therefore, musea need to focus on trust modeling and to evaluate trust levels of annotations before delivering them. However, because of the workload and specific skills required, it may result infeasible. For this reason we introduce a model which aims to automatically assess these trust levels.

As we saw in section 1.1, quality is a key feature within a professional media environment. Precision of the terms that are used plays a crucial role in determining the quality of annotations. Its achievement is mainly due to two factors: the annotator, which needs to be in possess of the necessary skills, and the thesauri or knowledge repository from which the information for annotating is chosen, which should assure a minimum level of reliability. Although a high reputation of the expert and of the source of information can be an important assurance about the correctness of the annotation, we still need to talk of trusted instead of correct annotations, since these evaluations are made by reasonably confident inference and not by a direct manual check and this implies the possibility that the annotation is not really correct.

2. RELATED WORK AND APPROACH

The problem of dealing with trust in semantic web annotations has already been addressed by various authors, and the notion that semantic web annotations have different levels of trust depending on their sources is a well-known consequence. Other approaches include the use of possibilistic logics [6] or belief [14, 16].

Our approach is to use RDF/OWL [17, 19] in association with subjective logic [10]. RDF/OWL is a family of languages that is commonly used for metadata management. Our approach aims to answer different needs at the same time. First of all, since we need to reason about metadata, that is, on RDF statements, we use RDFS [18] reification to reason about each element of such an annotation and to evaluate it in detail.

Secondly, since an estimate of the correctness of an annotation may be inferred by different metadata regarding it, which sometimes are orthogonal to each other, we choose to use Bayesian networks as a powerful reasoning tool to manage all these information sources. Heterogeneity of metadata has already been addressed by Damiani *et al.* in [6], and we choose to build our Bayesian network with subjective logic exactly because it allows to uniformly represent trust values, by collecting evidence and numerically evaluating it. On the other hand, heterogeneity of information sources is not only a characteristics of the sources themselves, but also of their distribution. For this reason, the logic we choose falls into a so-called "belief approach", like [14, 16] do. If Bayesian networks allow to infer a probability for a certain event, given the appearing of some related events, then the belief-based approach differs from a pure statistical approach in the fact that it quantifies uncertainty. Like the classical statistical approach, this also allows to compute the inferred probability for a given event, but the calculated probabilities take into account the amount of evidence collected so far. By contrast, a classical statistical approach considers only the ratio between considered events to calculate the probabilities.

Finally, an important requirement for our reasoning tool is the ability to merge contributions of different sources of information in order to obtain a final value that encloses all these contributions. Subjective logic offers a range of operators which, according to the existing relation between the information sources (*e.g.* dependence or independence), allows to merge the different contributions properly. The model focuses mainly on the evaluation of the trust level of the annotations made internally by the authority. We allow the model to collect external data, but only for ontological reasoning (*i.e.* to increase the availability of meta-information). Therefore, despite Richardson *et al.* [14], Golbeck [7] and Kamvar *et al.* [13], we don't focus on the distribution and collection of such evaluations. However, we will discuss the possibility of doing this in the future work section, 5.2.

2.1 Subjective logic

Trust is an activity that, by definition, involves probability. Indeed, if we decide to "trust" in someone or something, we do it because we (strongly) believe that, *e.g.*, this someone is really able to cope with what he or she promised to do, but we cannot be completely sure about it. We can receive or compute a lot of evidence about his or her trustworthiness but if we would be completely sure about it, we would simply know that particular property or action, we wouldn't need to trust in it.

Therefore, the use of a probabilistic approach for this kind of problem is natural. On the other hand, since evidence is gathered from multiple sources of information with different dependency relations to each other, we need to be able to represent these relations and exploit them. The most natural way to do so is to use Bayesian networks [9]. These are probabilistic graphical models that represent a set of random variables and their conditional dependencies via a directed acyclic graph. These random variables represent the sources of information about the evaluated annotation or the intermediate values obtained by combining them. These networks allow to combine the values assumed by the random variables, taking into account dependency and allowing us to obtain a unique final probability about the trustworthiness of the evaluated annotation. Subjective logic is a probabilistic logic which belongs to the so-called family of evidential reasoning tools and belief-based methods. The fact that it is classified as an evidential reasoning tool is clearly due to its ability to represent and manage evidence regarding particular events of interest. Moreover, subjective logic is a belief-based method, because the probabilities which it computes are estimated by reasoning on the quantity of evidence available. These probabilities therefore depend on the amount of evidence available: the more evidence we have, the more we "reasonably believe" in the ratio computed. For instance, 0.5 probability on two pieces of evidence is much less "believable" than 0.5 probability computed from a set of 200 pieces of evidence.

The key concept of subjective logic is the concept of opinion. An opinion given by a subject y about an assertion x is a quadruple $\omega_x^y(b, d, u, a)$, where $b, d, u, a \in [0...1] \wedge b+d+u = 1$. An opinion defines a degree of trustworthiness for the assertion x according to what y believes. When $b = 1$ or $d = 1$, then the opinion is equivalent to the boolean value of TRUE or FALSE. Otherwise, the fact that $b + d < 1$ and therefore $u > 0$ implies a certain degree of uncertainty in the opinion. The parameter a is the "a priori probability" (also called "base rate") about the outcome in case of absence of evidence. In absence of information that makes some possible outcomes more preferable than others, $a = 1/n$, where n is the number of possible outcomes.

An important characteristic of this method is the possibility to compute a probability both in the presence and in the absence of evidence. When more evidence becomes present, the computed probability becomes closer to the real probability. When evidence is present in a small amount, then the computed probability becomes close to the a priori probability a. This is given by the formula: $E = b + a * u$, where E is the expected value, which will also be referred to as "trust value". This formula shows how uncertainty modeling is similar to incorporating the margin of error into the probability calculation.

In our context, the possible outcomes are two: an annotation made by an actor is correct or not correct. Therefore we will consider a binomial distribution, that is a probability distribution with two possible outcomes, the first with probability p, the second with probability $1 - p$. In the absence of different indications, the possible outcomes will have the same probability, that is $a = 0.5$, because we cannot say anything a priori. In case, for instance, we know that the actor belongs to a trustworthy category the base rate would differ from 0.5, and the same would hold for any actor belonging to the same category. So, the base rate expresses the general expected a priori behavior for the actors belonging to a particular class of individuals. If evidence regarding a particular actor is present, that is evidence deriving from his history, then this yields a higher reliability when evaluating his ability. These are captured by the three b, d, u values, which are computed as follows:

$$b = \frac{p}{p + n + 2}$$

$$d = \frac{n}{p + n + 2}$$

$$u = \frac{2}{p + n + 2}$$

where p is the amount of positive evidence collected and n is the amount of negative evidence. These relations are derived from the so-called "Dirichlet probability distribution", on top of which subjective logic is built (see [10]). Intuitively, the number 2 in the denominators and the numerator of u is basically due to the fact that we have two possible outcomes (Accepted and Refused). In case this value would be higher than the number of possible outcomes, this would result in new observations having relatively less influence over the Dirichlet probability distribution, which means that the more recent evidence would contribute less than the older ones in the trust level computation (see [10]).

These formulas directly connect the belief value and the amount of evidence. From them we can infer that:

- the higher the amount of evidence, the lower the uncertainty;
- the lower the uncertainty, the lower the weight of the base rate in the computation of the expected value;
- the higher the amount of positive/negative evidence, the higher the belief/disbelief.

Belief values can also be represented within a triangular space, as depicted in Figure 1.

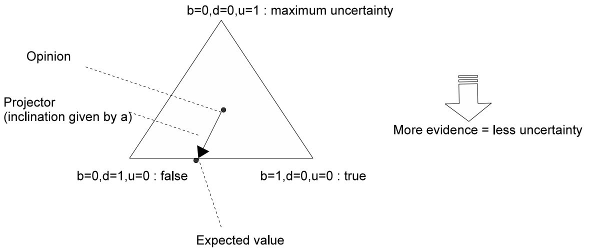

Figure 1: Graphical representation of a subjective logic opinion.

Our aim is to compute the probability that a certain author will produce correct annotations. In absence of information this probability is put at 0.5: we have neither reason to privilege the belief that he works correctly, nor the opposite. The more evidence (and, in general, information) we gather, the more we can restrict the variability of the possible outcomes and eventually move the expected value according to the distribution of the evidence collected.

An opinion represents a judgement given by a single point of view. The fact that the system is based on opinions represents also the main strength of the model, since they allow one to consider different points of view when evaluating annotations. Indeed, since we are not dealing with a direct content-based evaluation, and since available metadata could be different, orthogonal and at the same time all relevant, this method ideally can deal with all these points of view, combining them into a unique final result.

Another important feature of this approach is the possibility to build networks of information sources. Indeed, when analyzing an annotation, we could have different opinions originating from different sources of information, which may be independent or not. Since our aim is to obtain a final probability that the analyzed annotation is correct, we need to properly merge all these opinions in a final one, enclosing and mediating all these different contributions. Several operators are provided in subjective logic, such as the "consensus operator" [12]. According to the dependence between the opinions considered, these operators give, for instance, an average of the opinions weighted according to uncertainty, which means that the outcome opinion incorporates all input opinion, but giving more weight to the opinions based on the higher amount of evidence.

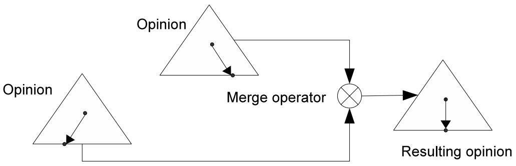

Figure 2: Network of opinions. Given the two initial opinions, the merge operator combines them in a final opinion which takes into account both points of view.

3. TRUST MODEL

We propose a model based on semantic web technology for the representation and ontological reasoning part, and subjective logic for the probabilistic reasoning part.

The aim of the model is to provide a tool for the automatic estimation and evaluation of trust levels of annotations. Our main purpose is, on the one hand, to reduce to a minimum the amount of human work necessary to make the model accurate (initial knowledge necessary to calibrate the model), and on the other hand, to increase accuracy as much as possible, in order to make the model itself trustworthy.

3.1 Data representation and ontological reasoning

Trust data is in fact a special form of metadata. As said before, we therefore use RDF/OWL for the representation of trust data. In particular, annotations are represented in RDF, and through RDFS we reify them in order to record metadata. Typical examples of metadata are the author of the annotation, which is linked to the reified annotation, or the author of a taxonomy used in the annotation, which is linked to the object of the annotation, in case of an annotation using taxonomies. When possible, we used standard ontologies like FOAF[1] and Dublin Core[2] to represent these metadata. However, we developed also a small ontology available online[3] in order to fully satisfy our requirements. For instance, we need to represent specific annotations which make use of taxonomies, and this implies the need both to represent meta-information regarding the taxonomy itself and to reason about the connection between the annotated object and the taxonomy elements (since, *e.g.* genus may be correct but species not, we avoid to treat the taxonomy as a unique entity).

Moreover, by exploiting Linked Open Data [1], we can enlarge the availability of metadata and, therefore, increase the number of possible sources of information about the trustworthiness of annotations. For example, if we consider the annotation of an artwork or an animal specimen, then meta-information about the term or taxonomy used to annotate could be limited when simply relying on data internal to the authority. With Linked Open Data we can gather information regarding the painter used to annotate the artifact or the taxonomy used to annotate the specimen. Using this additional evidence, we can more confidently check the correctness of the annotation.

3.2 Evidential reasoning

Once we have gathered enough semantically significant metadata, we can merge all contributions in order to obtain a single value representing the probability that the evaluated annotation is correct. As explained in section 2.1, subjective logic is the method we choose to solve this task.

Firstly, it allows us to uniformly quantify heterogeneous contributions. This is done by counting the number of positive and negative evidences attributed to each piece of metadata in the current context. Usually, positive and negative evidence is represented by samples of correct and wrong annotations evaluated by the authority.

Secondly, it allows us to merge all contributions in a convenient way. For instance, opinions deriving from different metadata could be dependent or independent of each other, or they could be more or less important within certain contexts. Through particular operators such as the "consensus operator" and the "discounting operator" [12], subjective logic can deal with all these relations between metadata and merge the contributions properly.

Finally the method is quite elastic, since it allows us to reason about the evaluated metadata in the absence of enough direct evidence, but in the presence of particular information about the average behavior of the category to which the metadata belongs. For instance, although we may not have information about a particular author, by knowing that he owns a particular diploma, and by knowing that on average people belonging to that "class" annotate correctly, the base rate associated to this annotator could be higher than 0.5. A higher base rate, implying the presence of some information, could, if this information is sufficiently reliable, allow us to take a decision.

Subjective logic can also be used to control the behavior of the system. By sampling and controlling the system's reliability, we can build an opinion about its reliability and then weight opinions on annotations according to these opinions. This can be seen as a web of trust, since by adding this layer, we build a reputation for the system that is returning us reputations about annotations.

3.3 Implementation

For data manipulation we use the SWI-Prolog semantic-web package[4]. We developed trust-management procedures in Prolog. The model has been developed according to the following structure:

Subjective logic module: This module has been developed as a generic subjective logic module and leaves aside any domain-related issues. This module, therefore, contains all the tools needed to represent subjective logic opinions, merge and discount them and

[1] `http://xmlns.com/foaf/0.1/`

[2] `http://purl.org/dc/elements/1.1/`

[3] `http://www.few.vu.nl/~dceolin/annotationTrust.rdf`

[4] `http://www.swi-prolog.org`

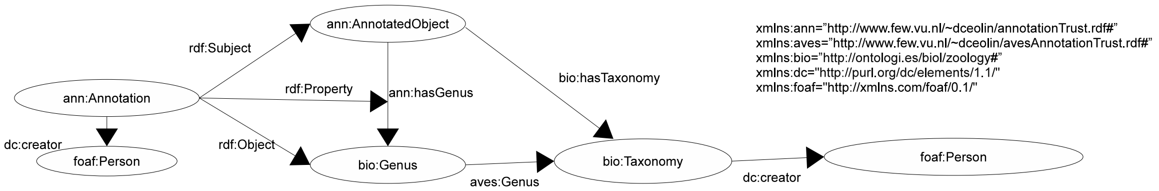
ann:AnnotatedObject
rdf:Subject
xmlns:ann="http://www.few.vu.nl/~dceolin/annotationTrust.rdf#"
xmlns:aves="http://www.few.vu.nl/~dceolin/avesAnnotationTrust.rdf#"
xmlns:bio="http://ontologi.es/biol/zoology#"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:foaf="http://xmlns.com/foaf/0.1/"
bio:hasTaxonomy
ann:Annotation
rdf:Property
ann:hasGenus
dc:creator
foaf:Person
rdf:Object
bio:Genus
aves:Genus
bio:Taxonomy
dc:creator
foaf:Person

Figure 3: Annotation representation with RDF. By the reification of the annotation, that is by the treatment of the annotation as an object, we can easily enrich it with the meta-information we collect.

record evidence. Moreover, it includes a set of predicates that allows different kinds of evidence management. For instance, these include the possibility to count all the evidence available, or to give more importance to the most recent ones, by giving less weight to the less recent, and the usage of a so-called sliding window. The sliding window allows one to take into consideration only the last pieces of evidence when evaluating a reputation. This module is available online[5].

Domain-related file: This file collects all the domain-related predicates. Here are defined the conditions for positive and negative evidence, as well as more generic strategies for evidence management and error handling. Finally, within this module the implementation choices are taken regarding evidence management, by choosing a suitable strategy among those offered by the subjective logic module.

3.4 Decision strategies

Our model aims at calculating the probability that a certain annotation is correct. Taking a decision always implies a degree of error, but errors may not always have the same importance. *E.g.*, within some contexts a false positive can be less desirable than a false negative. Different strategies are suitable for different application domains. We propose some decision strategies:

Fixed threshold: Once we have decided the maximum level of error acceptable, we will accept all annotations with a trust level above it. Clearly, there may be two kinds of threshold, one for acceptance and one for refusal. In the latter case, annotations are refused only when the trust level is below such a threshold. In case these different thresholds coexist, we have to take into account the fact that they outline a middle section, below the threshold of acceptance, but above the threshold of refusal, where our model is not able to evaluate annotations. For instance, we could decide that we can accept a maximum level of error of 10% due to acceptance of false positives. Therefore, we will accept all annotations with a trust level above 0.9. Since a trust level of 0.9 means that on average no more than one annotation out of ten is wrong (but we don't know which one), our false positive rate will be at most 10%.

Probability distribution simulation: The previous strategy guarantees a certain maximum error rate, but on the other hand, it does not leave room for improving it, since it accepts any annotation which trust level is beyond the threshold without trying to discover possible wrong annotations among those with the higher trust level (which may exist since their trust level is still lower than 1). In order to try to do this, we can try to simulate the probability distribution determined by the trust level and use such a simulation to take decisions. Suppose that an annotator has a reputation of 0.9. This means that, on average, he will make one wrong annotation out of ten. If our function accepts one annotation out of ten, when they are made by this author, then our error rate may reach 0%, in case we are able to match the wrong annotation with the refusal by the function, or diverge otherwise. By running this function multiple times, checking the expected value and variance of the results, and by trying to limit its deviation, we can at least infer useful information on which we can base our decisions.

Speed of variation Within certain domains, positive opinions coming from different sources "sustaining" each other may lead to a decision, although the final opinion resulting from their merge may be slightly different from 0 or 1. In particular, when we face an opinion which is positive or negative, but still far from acceptance or refusal and by merging it with another one regarding the same subject, our total opinion moves rapidly to one extremal value, this may be enough to take a decision. This strategy is frequently used when taking a decision based on a Bayesian network.

Another important aspect that has to be taken into account is the choice to reuse evaluations made by the model as evidence. On the one hand, this may be an optimal choice looking at the dependency of the data, since it allows to reinforce the strength of opinions without the need for more manual evaluations by the authority. On the other hand, this may also be a risky choice, since in case we make evaluations based on a not completely sure reputation, this increases the error rate.

These are the approaches analyzed so far, but clearly, this is not an exhaustive selection. However, the decision strategy prescinds from the calculation of the trust levels, which is the primary aim of the model, unless we don't reuse evaluations as evidence.

In the case study presented in section 4.2, we will use only the fixed threshold strategy.

3.5 Usage of the model

The model can uniformly deal with heterogeneous metadata about the annotations. This uniform representation of trust evaluation leads to two important consequences. The first is clearly the possibility to merge all these various contributions into a unique value.

[5] `http://www.few.vu.nl/~dceolin/subjective_logic.pl`

The second is reusability of this value. By clearly defining the context, the authority creating it, the metadata used and the methods applied, we facilitate their reuse. Another authority needing to evaluate the handiwork of the same author may directly make use of such evaluations, taking into account the reputation of the assessing authority and the methods used for the assessment. This way we implicitly allow the creation of a so-called "web of trust" [11].

4. CASE STUDY: NATURALIS DATA

The case study we face regards the annotation of bird specimens curated by the National Museum of Natural History in Leiden, Naturalis.

4.1 Data set

Naturalis Museum recorded in a database a series of annotations of bird specimens owned by it. This database records information about taxonomies, specimens, and how these are classified using taxonomies. Experts annotate each specimen using a taxonomy recorded in the database. The result of such a linking is a "one-to-many" relation, since in general more specimens of the same species are present. However, these annotations are not always correct, and this may be due to many reasons: for instance a mistake by the annotator, or the fact that the taxonomy became obsolete after a certain period. Therefore the museum also created a series of annotations, which it considers correct. Since the museum is the authority we refer to, this series of annotations is our landmark: our model should assign a high trust value to annotations produced by an annotator and confirmed by the museum, and a low trust value to the others. From a comparison between the trust values and the decision by the museum, we are therefore able to evaluate our model. For reasons of confidentiality, we cannot expose this dataset in full detail. At the "Netherlands Biodiversity Information Facility" portal[6], it is possible to see examples of correct annotations exposed by Naturalis Museum.

4.2 Case study setup

Data are provided in the form of a classical relational database. Through the use of D2RQ [3], these are easily converted into RDF. Once converted into RDF, we reify annotations, in order to associate also the creator with the annotations itself. The same process is done when enriching the taxonomy with additional information. Since taxonomy authors are recorded in a non-homogeneous way, we refer instead to the U.S. National Biological Information Infrastructure[7] to collect this kind of information. This infrastructure exposes an authoritative and exhaustive database of taxonomies which, once converted into RDF, has been used to annotate annotations we are evaluating. In order to improve the representation of reified birds annotations, we developed a small ontology available online[8], which extends the one cited in section 3.1 in order to accurately represent taxonomic annotations of bird specimens. To represent taxonomies, we use the Biological Taxonomy Vocabulary[9].

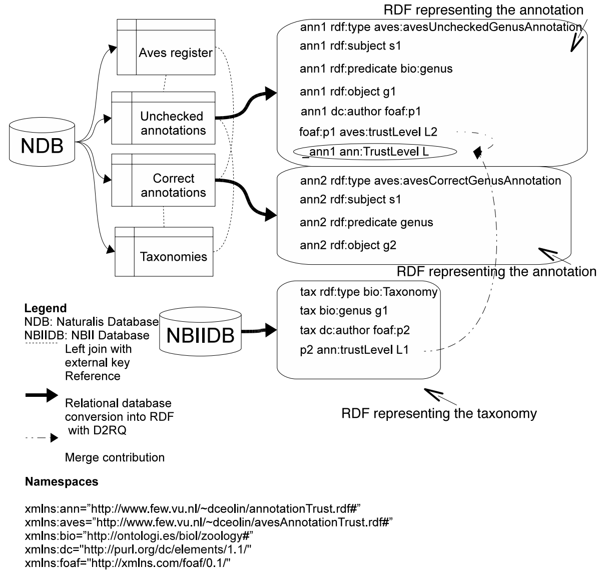
RDF representing the annotation
Aves register
Unchecked annotations
Correct annotations
Taxonomies
NDB
ann1 rdf:type aves:avesUncheckedGenusAnnotation
ann1 rdf:subject s1
ann1 rdf:predicate bio:genus
ann1 rdf:object g1
ann1 dc:author foaf:p1
foaf:p1 aves:trustLevel L2
ann1 ann:TrustLevel L
ann2 rdf:type aves:avesCorrectGenusAnnotation
ann2 rdf:subject s1
ann2 rdf:predicate genus
ann2 rdf:object g2
RDF representing the annotation
tax rdf:type bio:Taxonomy
tax bio:genus g1
tax dc:author foaf:p2
p2 ann:trustLevel L1
NBIIDB
RDF representing the taxonomy
Legend
NDB: Naturalis Database
NBIIDB: NBII Database
Left join with external key Reference
Relational database conversion into RDF with D2RQ
Merge contribution
Namespaces
xmlns:ann="http://www.few.vu.nl/~dceolin/annotationTrust.rdf#"
xmlns:aves="http://www.few.vu.nl/~dceolin/avesAnnotationTrust.rdf#"
xmlns:bio="http://ontologi.es/biol/zoology#"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:foaf="http://xmlns.com/foaf/0.1/"

Figure 4: Case study overview. Given the two initial databases, we create RDF representations of annotations and taxonomies. Afterwards, we merge all reputations available in order to obtain a unique trust value for the annotation.

Once the data had been prepared, we created a series of Prolog procedures, available online[10], which allowed us to build reputations for each kind of information source and then compute trust levels of annotations. Different implementation strategies have been adopted and the results are reported in section 4.3. The subjective logic predicates used by these procedures are those contained in the module described in section 3.3

4.3 Results and analysis

We analyzed a set of 65,600 annotations made by ten authors. We adopted different implementation strategies both to compare them and to simulate different scenarios. The results are presented in Table 1.

Nr.	Trainingset	Information sources	Error handling	Accuracy
1	30% Data	Author	No	43%
2	10 per source	Author	No	53%
3	10 per source	Author, Taxonomy	No	60%
4	10 per source	Author ∧ Taxonomy	No	76%
5	10 per source	Author ∧ Taxonomy	Yes	82%

Table 1: Results with different strategies.

Each strategy works on the same data, splitting them in a training set and a data set, but the way these subsets are built changes for each strategy: for instance, some strategies take the first 30% of data as trainingset, others consider a fixed amount of data for each information source. This, and

[6] http://www.nlbif.nl

[7] http://www.nbii.gov

[8] http://www.few.vu.nl/~dceolin/avesAnnotationTrust.rdf

[9] http://ontologi.es/biol/zoology

[10] http://www.few.vu.nl/~dceolin/naturalis.pl

the other differences explained in the following paragraphs, lead us to different results.

Strategy 1. The first solution adopted is the simplest one: fixed threshold (see section 3.4). We evaluate the reputation of each author considering the first 30% of annotations ordered according to the date of creation. This leads to a poor result, that is, an accuracy of evaluations of 43%, mainly due to two reasons. First, considering only one kind of metadata and a fixed threshold, then once it is established where the author is situated (above or under the threshold, that is accepted or not), there is no way to adjust his evaluation, since no different point of view is taken into account. Second, since authors are not uniformly distributed in the dataset (some authors started working on the dataset earlier, some later), we cannot gather enough evidence for all annotators. This leads to a conservative consequence: since we have no information to evaluate such annotations, these will always be refused (since false negatives are preferred to false positives), decreasing accuracy.

Strategy 2. The second solution solves one of the previous problems, that is the non-homogeneous distribution of the authors over the dataset, since it collects a fixed amount of evidence before using reputations to evaluate the annotation. This means that each reputation is used only after having collected a reasonable amount of evidence, and clearly this helps to improve the results. The improvement is quite significant but, although the performance is slightly better than what we would have tossing a coin, we are still far from a positive result.

Strategy 3. The third solution uses two sources of information, the reputation of the author of the annotation and the reputation of the author of the taxonomy. This second source of information is chosen because a typical reason for refusing this kind of annotation is the fact that the taxonomy used may have become out of date. By looking at the author of the taxonomy, we implicitly take into account the period when the taxonomy was created and the methods used for assessment, which are important indicators whether the taxonomy is out of date. Moreover it incorporates the previous improvement and takes a fixed amount of evidence for each source of information. These improvements lead us to an accuracy of 60% which, although far from an optimal result, again shows a substantial improvement. This solution is important because it shows how it is possible to successfully merge contributions from different sources in order to obtain a more precise result.

Strategy 4. The fourth solution reaches 76% of accuracy. This variant builds opinions based on the performance of each author with each taxonomy. Compared to the previous version, which took the two reputations and merged them, this is more precise, since it evaluates the contribution given by these reputations, taking also into account the existing relation between the subjects to which these reputations belong. So, when an author *a1* has a certain reputation, this is computed according to his behavior over time. The same can be said about the taxonomy *t1*. By analyzing the annotation made by *a1* using *t1*, in the previous strategy we merged their reputations, which were considered two distinct inputs. That approach is quite realistic, since it simulates the case when we collect opinions coming from different sources about different metadata of the same annotation. But using strategy 4 we can be more precise, by looking at the reputation of the author with a particular opinion, *i.e.* the reputation of $a1 \wedge t1$.

Strategy 5. The fifth solution gives the best result: 82% of accuracy. It starts from the improvement achieved with the previous strategy and adds an error handling procedure. This procedure monitors the behavior of the system and checks if annotations accepted by the model are really correct annotations and vice-versa. So, beyond evidence about authors of annotations and taxonomies, the procedure collects also this kind of evidence and, in case the precision goes below a certain threshold, then it firstly improves the reputation of the considered sources by collecting new evidence about them and secondly collects new evidence about the behavior of the system, in order to see if the more accurate reputations did actually improve the system behavior.

5. DISCUSSION

Due to the explorative nature of this work, we cannot derive any strong conclusions. However, we can suggest best practices which could help to reason about trust and to represent it. Further research will be needed to confirm these guidelines.

5.1 Best practices

From the Naturalis Museum case study, we see how our model could be supported and improved by the adoption of some best practices by the authority. These may include:

- The usage of RDF as standard language for metadata representation. Although any database can be easily “triplified” (see also [3] and [4]), since RDF is the standard technology for metadata representation, its usage is desirable.

- The usage of references (URIs) to standard knowledge repositories for annotations. Instead of building an internal knowledge repository for annotations, if possible, it is preferable to refer to repositories offered by authorities in the field. For instance, a taxonomy used to annotate may frequently be taken from standard authorities. In case of biological taxonomies, *e.g.*, the U.S. National Biological Information Infrastructure offers an authoritative and exhaustive database of known taxonomies. This helps to keep the meta-information about annotations consistent and uniform.

- Keep a log and profile for each annotator. From the profile, for instance, we can retrieve information that is useful to assess a precise a priori probability for an annotator’s ability to annotate.

- Record physical information about the annotated object. Any kind of evidence useful to assess the correctness of annotations should be recorded and evaluated. In particular, this kind of data can reveal a direct link between an annotated object and its annotation, by the coincidence of *e.g.* shape, color or dimensions of the object and, for instance, the species represented by the taxonomy.

5.2 Future work

As we can see from section 4.3, the model reasonably solves our trust evaluation problem, as well as the related problem of merging different heterogeneous contributions. Anyway, it leaves room for improvement, mainly in two distinct directions. The first direction regards directly the improvement of results. By redefining the features by which we infer the trust values, by increasing the amount of features considered, and by connecting to more external sources to get more metadata regarding our annotations, we can improve our estimates without changing the data. Moreover, other techniques could be investigated in order to analyze metadata more effectively. The second direction regards the possibility to make these estimates available over the web. Moreover, from an annotation's trust value we can estimate the author's expertise about the annotation's topic. One starting point for this is the hoonoh ontology for expertise representation [8], which allows one to represent expertise by linking people with subjects with weighted links. Anyway, as we saw in section 3.5, we would need to extend such an ontology by representing the authority which assessed such a value, the method used and the evidence considered, in order to decentralize our trust management model (see also [15]). Moreover, this opens a privacy issue, since we should take into account the willingness of the evaluated expert to publish these data. By solving this, we could also address problems of confidentiality of the kind we had with the full exposure of our dataset, since privacy-related consequences may reduce the authority's willingness to expose data.

6. ACKNOWLEDGEMENT

Heartfelt thanks to the National Museum of Natural History in Leiden, Naturalis for allowing the realization of this work by sharing the data used in the case study.

7. REFERENCES

[1] Linked data - Connect Distributed Data across the Web, Mar. 2010. http://linkeddata.org/.

[2] Steve Museum, Mar. 2010. http://steve.museum/.

[3] The D2RQ Plattform - Treating Non-RDF Databases as Virtual RDF Graphs, Mar. 2010. http://www4.wiwiss.fu-berlin.de/bizer/d2rq/.

[4] Triplify.org, Mar. 2010. http://triplify.org/.

[5] L. J. Camp. Designing for trust. In R. Falcone, K. S. Barber, L. Korba, and M. P. Singh, editors, *Trust, Reputation, and Security: Theories and Practice, AAMAS 2002 International Workshop, Bologna, Italy, July 15, 2002, Selected and Invited Papers*, volume 2631 of *Lecture Notes in Computer Science*, pages 15–29. Springer, 2002.

[6] P. Ceravolo, E. Damiani, and C. Fugazza. Trustworthiness-related uncertainty of semantic web-style metadata: A possibilistic approach. In F. Bobillo, P. C. G. da Costa, C. d'Amato, N. Fanizzi, F. Fung, T. Lukasiewicz, T. Martin, M. Nickles, Y. Peng, M. Pool, P. Smrz, and P. Vojtás, editors, *Proceedings of the Third ISWC Workshop on Uncertainty Reasoning for the Semantic Web Busan, Korea, November 12, 2007*, volume 327 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.

[7] J. Golbeck. Combining provenance with trust in social networks for semantic web content filtering. In L. Moreau and I. T. Foster, editors, *Provenance and Annotation of Data, International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006, Revised Selected Papers*, volume 4145 of *Lecture Notes in Computer Science*, pages 101–108. Springer, 2006.

[8] T. Heath and E. Motta. The hoonoh ontology for describing trust relationships in information seeking. In *Personal Identification and Collaborations: Knowledge Mediation and Extraction (PICKME2008)*, 2008.

[9] D. Heckerman. A tutorial on learning with bayesian networks. Technical report, Learning in Graphical Models, 1996.

[10] A. Jøsang. Probabilistic logic under uncertainty. In J. Gudmundsson and C. B. Jay, editors, *Theory of Computing 2007. Proceedings of the Thirteenth Computing: The Australasian Theory Symposium (CATS2007). January 30 - Febuary 2, 2007, Ballarat, Victoria, Australia, Proceedings*, volume 65 of *CRPIT*, pages 101–110. Australian Computer Society, 2007.

[11] A. Jøsang and T. Bhuiyan. Optimal trust network analysis with subjective logic. In *Proceedings of the Second International Conference on Emerging Security Information, Systems and Technologies, SECURWARE 2008, August 25-31, 2008, Cap Esterel, France*, pages 179–184. IEEE, 2008.

[12] A. Jøsang and D. McAnally. Multiplication and comultiplication of beliefs. *Int. J. Approx. Reasoning*, 38(1):19–51, 2005.

[13] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 640–651, New York, NY, USA, 2003. ACM.

[14] M. Richardson, R. Agrawal, and P. Domingos. Trust management for the semantic web. In D. Fensel, K. P. Sycara, and J. Mylopoulos, editors, *The Semantic Web - ISWC 2003, Second International Semantic Web Conference, Sanibel Island, FL, USA, October 20-23, 2003, Proceedings*, volume 2870 of *Lecture Notes in Computer Science*, pages 351–368. Springer, 2003.

[15] F. Spiessens, J. den Hartog, and S. Etalle. Know what you trust. In J. D. G. Pierpaolo Degano and F. Martinelli, editors, *Formal Aspects in Security and Trust, 5th International Workshop, FAST 2008, Malaga, Spain, October 9-10, 2008, Revised Selected Papers*, volume 5491 of *Lecture Notes in Computer Science*, pages 129–142. Springer, 2008.

[16] L.-H. Vu and K. Aberer. Effective usage of computational trust models in rational environments. In *2008 IEEE / WIC / ACM International Conference on Web Intelligence, WI 2008, 9-12 December 2008, Sydney, NSW, Australia, Main Conference Proceedings*, pages 583–586. IEEE, 2008.

[17] W3C. Owl Web Ontology Language Overview, Mar. 2010. http://www.w3.org/TR/owl-features/.

[18] W3C. RDF Vocabulary Description Language 1.0: RDF Schema, Mar. 2010. http://www.w3.org/TR/rdf-schema/.

[19] W3C. Resource Description Framework (RDF), Mar. 2010. http://www.w3.org/RDF/.