Building Time-Affordable Cultural Ontologies Using an Emic Approach

Jean Petit, Jean-Charles Boisson, Francis Rousseaux

Building Time-Affordable Cultural Ontologies using an Emic Approach

Jean Petit[1,2], Jean-Charles Boisson[3], and Francis Rousseaux[2]

[1] Capgemini Technology Services, 7 rue Frederic Clavel, 92287 Suresnes
jean.petit@capgemini.com
[2] MODECO team, CReSTIC laboratory (EA 3804), University of Reims Champagne-Ardenne, France
francis.rousseaux@univ-reims.fr
[3] CASH team, CReSTIC laboratory (EA 3804), University of Reims Champagne-Ardenne, France
jean-charles.boisson@univ-reims.fr

Abstract. Recently, studies about culturally-aware systems have arisen to address digitized culture. Among these systems those enculturated driven by cultural knowledge embed culture in their design. To deal with the specifics of cultural groups, the development of machine-readable cultural knowledge representations can provide a substantial help. In this research we present a process to build time-affordable, emic, conceptually-sound and machine-readable cultural representations. These representations originate from Cognitive Anthropology. They follow a three steps methodology: ethnographic sampling, individuals' personal knowledge elicitation and cultural consensus analysis. We use lexico-semantic relation extraction as a mean to automatically elicit knowledge structures. Their formalisation is achieved through Ontology Engineering.
We conducted experiments to build three cultural ontologies in order to assess the whole process. It came out that with the lexico-semantic relation extraction technique, the best representations we can obtain are consensually-limited, incomplete and contain some errors. However, many clues indicate that these problems should be solved by using higher quality elicitation techniques.

Keywords: Cultural ontology, cultural representation, emic approach, culturally-aware systems, lexico-semantic relation extraction

1 Introduction

Interest in cultural awareness grows more popular as globalisation is vector of increasing cultural diversity. Since the 2000s, with the rapidly expending web, culture is digitized and computer systems are now the entities which are the most exposed to its diversity. Culture shapes users' behaviors and thus impacts the performance of many systems/applications. That is why these systems have to develop cultural awareness.

Blanchard et al. [1] define culturally-aware systems as "any system where culture-related information has had some impact on its design, runtime or internal processes, structures, and/or objectives". They present three types of systems: enculturated systems, runtime cultural adaptation systems and cultural data management systems. Enculturated systems are systems whose design meet the cultural requirements of given cultures [1]. Runtime cultural adaptation systems aim to artificially reproduce cultural intelligence through two steps: understanding and adaptation. In other words, by identifying one's culture a culturally-intelligent system can provide the right enculturation as presented by Rehm [2]. The enculturation of a system is constrained by the cultural knowledge available for the latter or a designer. That is why, machine-readable representations providing understanding about cultures could effectively support the development of these systems.

Two approaches can be used to produce representations of cultures. The etic approach has for objective to find cultural universals. It is an outsider view of culture. In contrast, the emic approach tries to identify the specifics of a culture such as their concepts and behaviors. Insight is gained from inside. Currently, cultural knowledge representations used to support the development of enculturated systems are etic-based. Their main appeals are that they are ready-to-use representations easily applicable to any culture [3, 4]. However, these representations are coarse-grain and limit the understanding of the cultures they describe [5]. Therefore, finer-grained emic-based representations are more relevant to develop enculturated systems.

While emic-based representations solve the problem associated with the lack of granularity, their creation is time-consuming. Most of the methodologies used in practice by ethnographers require intensive human intervention (from the ethnographers or participants) in the process of eliciting knowledge. Therefore the latter is hardly scalable, and thus not practicable to deal with the diversity of cultures. As such, the process supporting the construction of emic-based cultural representations must be relatively automatic.

In this paper we present a process applicable to any cultural domains to build time-affordable, emic, conceptually-sound and machine-readable cultural knowledge representations. To construct these representations we followed a methodology coming from Cognitive Anthropology. It is composed of three steps leading to the acquisition of culturally-relevant information: ethnographic sampling, individuals' personal knowledge elicitation and cultural consensus analysis. The time-affordable elicitation of knowledge and its formalisation are similar to what already exist in other ontology engineering works such as SPRAT [6] or DYNAMO[4] [7]. We follow Hearst's [8] method to automatically extract hypernym/hyponym relations from texts. As for the formalisation of the representations, we rely on the Resource Description Framework (RDF) formal language. Therefore, this research is about the emic and automatic generation of cultural ontologies from texts.

[4] *https://www.irit.fr/dynamo/*

Our plan is as follow. We begin by introducing the methodology. It starts with the creation of the cultural knowledge representations and ends with their formalisation. Then, we present our process and the associated design choices. We end by experimenting extensively our process on the public safety domain with police forces coming from Australia, USA and England. Obtaining encouraging results, we conclude this study.

2 Emic-based Cultural Knowledge Representations

Ethnography is the process of collecting, recording and searching for pattern to describe a culture of people. In other words, ethnography is about discovering cultural knowledge leading to the production of cultural knowledge representations. "New ethnography", ethnoscience or Cognitive Anthropology are founded on the premise that culture is a "conceptual mode underlying human behavior" [9]. The cognitive theory of culture situates culture in the mind as a system of learnt and shared knowledge [10, 11].

This theory shaped a number of methodologies to produce cultural representations which are intrinsically emic. "Ethnographers must discover the organizing principles of a culture–the semantic world of the natives–while avoiding the imposition of their own semantic categories on what they perceive" [12].

To our knowledge, there is no clearly defined methodology to create cultural representations. Most of the ones developed in the literature are based on the ethnographers' experiences. However, these methodologies share three main steps: ethnographic sampling, individuals' personal knowledge elicitation and cultural consensus analysis [13–16].

2.1 Ethnographic Sampling

The ethnographic sampling step is based on the idea that cultural knowledge is socially-constructed. It aims to capture a representative number of individuals likely to share the same culture and thus similar knowledge. This task is generally achieved through the identification of a community, a set of individuals with long-term, strong, direct, intense, frequent and positive relations [17].

Once the ethnographic sample is determined, the knowledge of each participant needs to be elicited.

2.2 Individuals' Personal Knowledge Elicitation

Knowledge is personal [18]. It roots deeply in the subconscious of one self in a tacit state. In order to elicit knowledge, it has to become object of thought [19]. The goal of the knowledge elicitation step is to explicit tacit internal knowledge structures. Jones et al. [20] distinguish two categories of knowledge elicitation: direct and indirect. In the first category, knowledge is directly elicited by the individual possessing the knowledge whereas in the second, knowledge emerges from the analysis of data collected from the individual.

"[C]oncepts are the building blocks of knowledge [and] relations [...] the cement that links up concepts into knowledge structures" [21]. Lexico-semantic relations are universal/intercultural knowledge structures representative of basic cognitive functions [21, 22]. They constitute the core of any conceptualisation. As such, individuals' knowledge elicitation is mainly about acquiring concepts and lexico-semantic relations.

After eliciting the personal knowledge structures of each individual constituting the sample, their distribution has to be analysed to determine their cultural dimension.

2.3 Cultural Consensus Analysis

The cultural consensus analysis step enables the operationalization of culture [15]. Cultural Consensus Theory (CCT) "formalizes the insight that agreement among [individuals] is a function of the extent to which each knows the culturally defined 'truth' " [23]. CCT also "refers to a family of models that enable researchers to learn about [individuals'] shared cultural knowledge" [24] such as the General Condorcet Model [25]. Depending on the form of the elicited knowledge, either formal or informal CCT models are used [26, 27]. However, simple aggregations, majority or averaging responses across respondents also constitute reasonable cultural estimates [28].

The three steps of the methodology leads to the production of cultural knowledge representations. However as such, they cannot be used for the development of enculturated systems as computers systems are not yet able to make sense of them. To be understandable, they have to be formalised.

3 Formal Cultural Knowledge Representations

The cultural representations are composed of knowledge structures. The formalisation of such structures is studied in the field of Knowledge Engineering, more precisely the Ontology Engineering subfield. Therefore, methodologies to build ontologies could be used to formalise the cultural representations.

3.1 Ontologies

Gruber had defined an ontology as "an explicit specification of a conceptualisation" [29]. The term 'explicit' in Gruber's definition means that the knowledge must be specified unambiguously, constraining its interpretation. The principal components of an ontology are labels, concepts, relations and axioms. Axioms are rules associated to the relations in order to embed logic necessary for reasoning.

Borst [30] added to the former definition that the specification had to be formal and the conceptualisation shared. Indeed, it is necessary that the conceptualisation results from a consensual agreement to ascertain that the knowledge embedded is coherent and consistent within a specific context. This task is called an 'ontological commitment'. This aspect is ensured by the shared dimension of

the cultural representations. The formalisation of the specification is needed for interoperability, re-usability and especially for for enculturated systems to read cultural representations.

There are different levels of formalism depending on the language used to express the ontology ranging from informal, mostly written in natural languages, to formal, based on machine-readable languages. Formal languages like RDF (Resource Description Framework) or OWL (Web Ontology Language) are supporting the semantic web. RDF is a language based on entities (*resource*, *property*, *value*) which constitute triples of the form (*subject*, *predicate*, *object*). Resources are concepts described thanks to an Uniform Resource Identifier (URI). It makes sense since ontologies are non-ambiguous specifications. Properties can be attributes or any other kind of relations, most likely semantic ones. Values are literals pointing either to a symbol or another resource. The common syntax to formalize RDF is the XML, called RDF/XML. Ontologies written in RDF can be interpreted by machines through SPARQL Protocol and RDF Query Language (SPARQL).

3.2 METHONTOLOGY

Methodologies to create ontologies are mostly based on experience [31]. The METHONTOLOGY is a proven framework describing the general steps to build an ontology [32]. Common steps are composed of specification, conceptualisation, formalisation, implementation and evaluation.

The specification consists in planning the production and exploitation of an ontology. At a minimum, it defines its primary purpose, level, granularity and scope. These specifications are mainly guidelines for the conceptualisation. Typically, the conceptualisation step is carried out by a group of domain experts. The goal is to discover the significant concepts and associated relations related to a domain [33]. The formalisation step expresses the conceptualisation with formal languages. It is often manually supervised by knowledge engineers or with the support of a software like Protégé[5]. Mapping techniques can also be used to automatically transpose informal to formal knowledge [34]. The implementation step addresses the technical and practicable aspects associated with the usage of an ontology by a computer system. The evaluation step validates each step according to the specifications.

Following the METHONTOLOGY, we are able to produce formal cultural ontologies by considering cultural knowledge representations as conceptualisations. Finally, these ontologies are readable by computer systems and can provide a significant amount of understanding about the cultures they represent.

[5] Available at: http://protege.stanford.edu/

4 Building Time-Affordable, Formal and Emic-based Cultural Representations

The design of our process was driven by the METHONTOLOGY whose conceptualisation step consists in the methodology coming from Cognitive Anthropology. Among other choices required to build the process, we decided to use the lexico-semantic relation extraction to have an elicitation as automatic as possible.

4.1 Selecting Individuals based on Shared Social Criteria

Typically, cognitive anthropologists select their sample through shared socially-related criteria such as genders, religions, jobs or areas - working places [35], towns [36] or regions [16]. While the strength of this method comes from its ease of use and speed, its weakness is that it cannot fully guarantee that the selected individuals actually represent a community. Effective but costly techniques to identify communities can be found in social sciences such as the community detection algorithms coming from social network analysis.

In this study, samplings are created by following the traditional technique as a number of studies proved its efficiency.

4.2 Eliciting Automatically Individuals' Knowledge from Texts

Automatically extracting individuals' knowledge structures from texts is an indirect elicitation technique [37]. It is composed of two tasks. The first one consists in collecting a sufficient amount of textual data for a given individual. The second task aims to retrieve the latter's knowledge (i.e. significant concepts and/or relations) by analyzing the data.

4.2.1 Collecting Web Data

Ethnographic data are mainly textual and most of the time collected thanks to interviews or observations. Besides being costly in time, recording data through these means also biases to some extent the data. The safest and fastest technique to collect data is to gather already existing raw data.

Nowadays, the web provides a large amount of freely available textual data about many individuals from which data can be collected. In our process, the data were retrieved directly from websites. Textual data collection was achieved thanks to HTTRACK[6]. It is a tool that can mirror the content of a website by crawling and downloading its files.

The automation of the data collection came with an additional constraint during the sampling step. Indeed, it became necessary to verify that the individuals composing the sampling disposed of accessible online data.

[6] *http://www.httrack.com/*

4.2.2 Textual Data Analysis

The goal of the data analysis is to retrieve the conceptualisation of an individual [37]. This part of our process consists in acquiring knowledge structures by mining significant concepts and their relations. It required several preprocessing steps. We started by cleaning the data, followed with natural language processing and ended by annotating the lexico-semantic relations to extract.

Preprocessing

The web nature of the data collected drove the cleaning operations. Web data can come in various file formats (.doc, .odt, etc). The text extraction from any files was achieved by Apache Tika[7]. We handled language heterogeneity by identifying the language of each document with the LangDetect API [38]. We only kept English documents. OpenNLP[8] was used to detect sentences. We decided to work on the sentence level rather than the document level mainly to avoid data redundancy by ensuring that the sentences were unique. For example, documents coming from websites are often distinct from each other while they are composed of duplicate contents such as menus, Twitter or Facebook feeds and so on.

Then, we used the Stanford CoreNLP API[9] to support common natural language processing operations: tokenization, Part of Speech (PoS) tagging and lemmatization. Eventually, nominals which constitute the main concepts of conceptualizations were found using a simple pattern matching based on the PoS tags of the tokens.

After having cleaned and preprocessed the textual data, the results were stored as annotations in a 'serial data store' using GATE[10] (General Architecture for Text Engineering). This last operation was required to easily retrieve and mine the data.

Discovering Important Concepts

Finding significant concepts in content is based on the idea that the number of occurrence and importance of a token are correlated. Thus, term frequency is often used to weight and rank terms. Other metrics can achieve similar results, such as TF/IDF (Term Frequency/Inverse Document Frequency).

In our process, the important concepts were selected by coupling the quantification of nominals with a rough filtering on their total occurrences.

Finding Significant Relations

In this study, we use the most popular method to find lexico-semantic relations. Introduced by Hearst [8], it relies on handwritten syntactic patterns indicative

[7] *https://tika.apache.org/*
[8] *https://opennlp.apache.org/*
[9] *http://stanfordnlp.github.io/CoreNLP/*
[10] *https://gate.ac.uk/*

of semantic relations. For example, in the sentence: 'A dog is an animal', the syntactic pattern 'is a' indicates that there is a hypernym/hyponym relation between 'animal' and 'dog'. Therefore, hypernym/hyponym relations can be discovered through a simple mapping, by using the expression Y *is a* X, with Y and X two nominals. Thereafter, many researchers confirmed the relevance of Hearst's methodology by applying it for other lexico-semantic relations [39–44].

Like Wang et al. [45], the implementation of lexico-semantic relation extraction was achieved through the Java Annotation Patterns Engine[11] (JAPE) which is specific to GATE. The syntactic patterns we used are summarized in the table 1.

Syntactic patterns
NP_1 *such as* NP_2
NP_1 *like* NP_2
NP_2 *and|or other* NP_1
NP_1 *for example|instance* NP_2
NP_1 *especially* NP_2

Table 1. Syntactic patterns indicative of hypernym/hyponym relations.

The final set of extracted lexico-semantic relations is constituted by filtering them according to the significance of their pairs of concepts.

At the level of the individuals, we are able to elicit their personal knowledge. However, we cannot yet determine which part is cultural. To this end, we have to analyze the 'sharedness' of these distributed knowledge structures.

4.3 Aggregating Concepts and Lexico-Semantic Relations

To analyze the cultural consensus of the sample, the elicited personal knowledge (concepts and lexico-semantic relations) of each individual was aggregated. It led to a mixed representation composed of knowledge ranging from personal to cultural (similarly to Vuillot et al. [16]). To obtain a valid cultural representation, it is necessary to evaluate the knowledge and filter the latter based on its distribution.

At this stage we are able to create a cultural representation from an ethnographic sample. However, these representations cannot be implemented into enculturated systems and thus are still unusable. They have to be formalised.

4.4 Ontologizing Concepts and Lexico-Semantic Relations

In our process, we used the "ontologizing" technique [34]. After the consensus analysis, we mapped the concepts and hypernym/hyponym relations into RDF classes and RDFs sub-classes.

[11] https://gate.ac.uk/sale/tao/splitch8.html

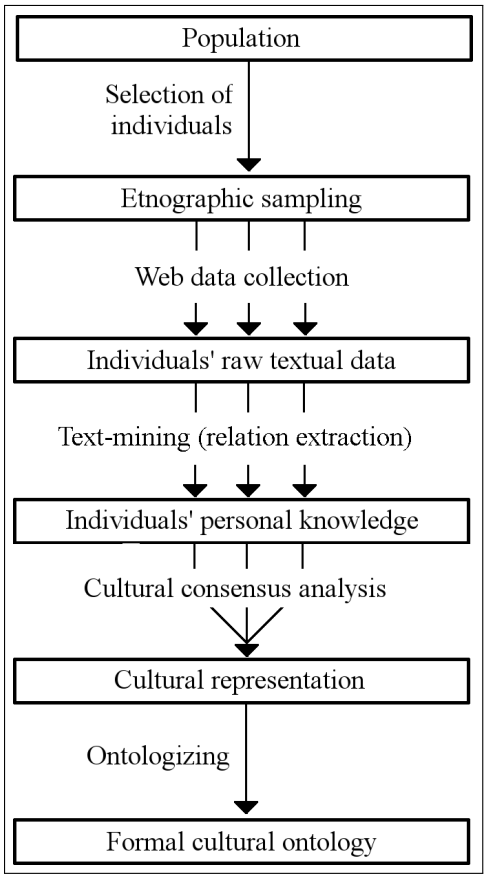

Fig. 1. Overview of the whole process to produce a formal cultural ontology.

The formalisation constitutes the last step of our process which is summarized in figure 1. It starts by selecting individuals based on shared social criteria. Then, web data about each individual of the sample are collected. These data are analysed through text-mining techniques to automatically elicit their respective personal knowledge (embodied in the conceptual structures). By quantifying the sharedness of individuals' personal knowledge, we are able to determine the cultural consensus. The latter analysis enables the production of a cultural representation. Finally by ontologizing the conceptual structures, a formal cultural ontology is created. Having described the whole process to produce formal time-affordable cultural representations, the next section consists of experiments to assess its performances.

5 Experiments

The public safety domain was chosen for our experiments for two main reasons: the available amount of data and current social context. After a description of the settings, we present and discuss the results associated to three formal cultural representations we tried to produce.

5.1 Settings

We constituted three samples with culturally different police forces coming respectively from Australia, United States and England (see table 2[12]). Considering agencies as individuals may not be the best choice to carry out our experiments. However, this decision was driven by the necessity of being able to collect large amount of textual data about a single domain for a consequent number of 'individuals'.

Sample	**Number of Individuals**
Australian Police Forces	7
American State Police Forces	21
English Police Forces	39

Table 2. Samples with their respective number of individuals.

While collecting data from the web, due to the robot protection or other factors, the content of some websites could not be retrieved. Therefore, we excluded these police forces from our samples.

After having retrieved the data, we preprocessed it. We cleaned the textual data and kept well formed sentences with a length between 40 and 500 characters. We removed police forces having less than 10,000 sentences left. This threshold was used to separate the individuals which possess too few data. The table 3[13] provides updated information about our samples.

Sample	**Individuals**	Minimum	Maximum	Mean
Australian Police Forces	5	28,303	57,583	47,988
American State Police Forces	10	21,256	174,314	76,161
English Police Forces	19	10,499	89,825	40,508

Table 3. Samples with the final number of individuals as well as the minimum, average and maximum number of sentences.

[12] The 'Appendix A' provides the detailed list of individuals we had at the initial stage of the process and their respective sample.

[13] Details about the police forces remaining and the associated number of sentences are available in 'Appendix B'.

Then, we quantified the nominals and extracted the lexico-semantic relations for each individual. For each sample, the nominals were ranked given their averaging position. We kept arbitrarily the top 1000 nominals and filtered accordingly the hypernym/hyponym relation candidates in order to create the various domain conceptualizations. At this point, we were able to produce cultural representations for the Australian, American and English police forces.

5.2 Evaluation

The evaluation of our experimental results was achieved by relying on a semi-automatically constituted gold standard. Three gold standards were constituted with labeled lexico-semantic relations, one for each sample. Because every police forces belong to the westerner culture, we were able to use WordNet [46], which possesses a similar cultural bias, to obtain automatically assessments on the elicited lexico-semantic relations. Then, we reviewed these relations to ensure their correctness as well as to validate contextual relations. Contextual relations are often considered as wrong [8], but for us they are relevant manifestations of cultural features, thus they were kept. For instance, we validated the relation (*issue*, *hypernym*, *crime*) or (*partner*, *hypernym*, *school*) because crime is an issue for police forces and English ones have often partnerships with schools. The raw results for each sample are given in table 4. It has to be understood that they are not based on consensus, thus not representative of cultural representations. They are produced with the mixed elicited knowledge of every individuals.

Sample	**Valid relations**	**Total relations**	**Precision**
Austrian Police Forces	272	884	30%
American State Police Forces	403	1290	31%
English Police Forces	647	2288	28%

Table 4. Raw results for each sample.

The precision of the extraction of lexico-semantic relation candidates is known to be relatively low. For the hypernym/hyponym relations, Cederberg and Widdows reported 40% [43], Maynard et al. 48.5% [6] and Hearst 52% [47]. Whereas, our raw results show an average precision of 30%. According to Cederberg and Widdows, the discrepancy in precision is mainly due to the difference of quality between the datasets. In fact, Hearst use Grolier's Encyclopedia, Maynard et al. use Wikipedia and themselves the British National Corpus. In contrast, we are using sources of poorer quality as our data came directly from website pages. We believe that it can explain our lower initial precision.

We observed the potential cultural representations by varying the number of agreements increasingly. We expected that highly consensual representations have higher precision but a lower relation coverage compared to mixed ones. Our hypothesis was that to obtain the best cultural representations, it is necessary to manage properly this trade-off between precision and loss. We computed the loss

as follow: $loss(n) = (v_1 - v_n)/v_1$, with n the minimum number of agreements ($n > 1$) and v the number of valid relations remaining. The new results are provided in table 5.

N.A.	**Loss**			**Precision**		
	A.P.F.	A.S.P.F.	E.P.F.	A.P.F.	A.S.P.F.	E.P.F.
2	76%	77%	61%	68%	61%	51%
3	93%	92%	79%	79%	79%	60%
4	98%	97%	87%	100%	92%	66%
5	99%	98%	91%	100%	100%	69%
6	-	99%	94%	-	100%	79%
7	-	99%	96%	-	100%	88%
8	-	99%	97%	-	100%	87%
9	-	-	98%	-	-	90%
10	-	-	99%	-	-	100%
11	-	-	99%	-	-	100%
12	-	-	99%	-	-	100%
13	-	-	99%	-	-	100%
14	-	-	99%	-	-	100%
15	-	-	99%	-	-	100%
16	-	-	99%	-	-	100%
17	-	-	99%	-	-	100%
18	-	-	99%	-	-	100%
19	-	-	99%	-	-	100%

Table 5. Loss and precision for each sample – Australian Police Forces (A.P.F.), American State Police Forces (A.S.P.F.) and English Police Forces (E.P.F.) – according to the number of agreements (N.A.).

Our first observation concerns the cultural dimension of our study. To produce cultural representations based on consensually shared knowledge, a weak agreement of at least half the sample is expected. Obtaining such a number in our experiment leads to cultural representations with a loss of 98% to 99% for a 100% precision. Such representations would have too few relations to be directly usable.

The second observation is that to obtain a satisfying precision (superior or equals to 90%), the loss is again too important: 98% for the Australian Police Forces, 97% for the American State Police Forces and 98% for the English Police Forces. The best trade-off is around 77% loss for 63% precision.

Our third observation is related to the practical aspect regarding the time required to produce cultural representations. To carry out the whole experiment, it took one full day using a normal laptop (by multi-threading it on a quad core computer with 16Gb rams). Using industrial means for production would shorten the necessary time in terms of minutes, thus leading to highly time-affordable representations. The problem is that based on the trade-off, reviewing the cultural representations for corrections will take hours or days.

Based on these observations, we conclude that the main problem is the high loss. The loss could be explained by three factors. The first one concerns the high number of relations specific to individuals such as (*partner*, *hypernym*, *northumbria police*), but it does not constitute a problem as we are not interested by those. The second factor corresponds to the cultural domain. Many extracted relations are related but do not strictly belong to the public safety domain like (*resource*, *hypernym*, *goods*). Similarly to the first factor, this loss does not matter. The third factor concerns the scarcity of the syntactic patterns enabling the extraction of the lexico-semantic relations. Their low recall has for consequence that the discovery of a relation in a corpora is related to luck. This last factor is truly problematic.

This issue is directly linked to the knowledge elicitation technique used in our study. Indeed, lexico-semantic relation extraction relying on syntactic patterns cannot provide the quantity nor the quality required to support properly our process to produce cultural representations. In fact, no existing hypernym/hyponym relation mining technique using large corpora might be able to achieve this task. So we were expecting those results.

Nevertheless, with few efforts we were able to produce a relevant partial cultural ontology for the English Police Forces composed of 131 hypernym/hyponym relations. We used Gephi[14] to visualize the end result.

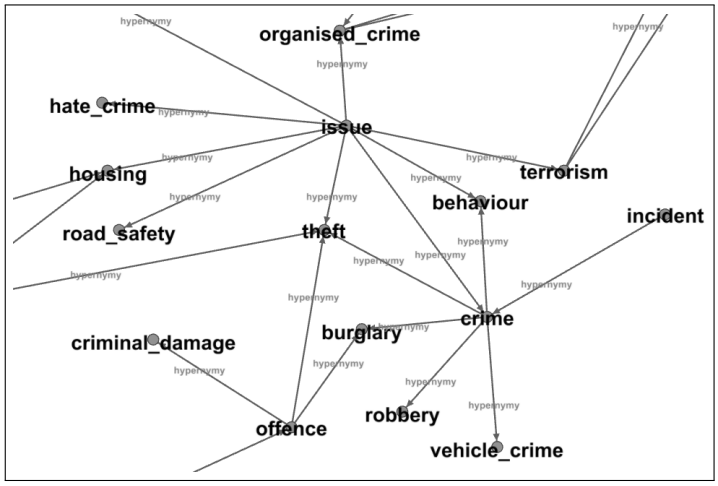

Fig. 2. Piece of the cultural ontology produced for the English Police Forces.

On figure 2 we focused on the concept 'crime'. We observe common hypernym/hyponym relations as well as an interesting contextual relation between 'hate crime' and 'issue'. Such relations are really meaningful in a cultural con-

[14] https://gephi.org/

text. In fact, the focus on hate crimes by English police forces comes from the enforcement of the Equality Act 2010[15]. It also becomes obvious that many relations are missing, but we believe that this representation provides a coherent foundation to support further improvements.

6 Conclusion

We have to remind that our goal was to build time-affordable, emic, conceptually-sound and machine-readable cultural representations. We introduced a methodology coming from Cognitive Anthropology to build emic-based cultural conceptualisations. In addition, we explained their formalisation through Ontology Engineering. Then, we presented a process to produce mostly automatically the representations. Using lexico-semantic relation extraction, the best we can obtain with this technique are representations consensually-limited, incomplete and containing some errors. However in the future, by using higher quality elicitation techniques, these problems could be solved.

Up to day, culturally-intelligent systems are developed using etic-based cultural representations. While facilitating cross-cultural mediation, these coarse-grain representations are not fitted for the development of systems requiring a deep understanding of cultural aspects [5]. We believe that the production of fine-grain cultural ontologies, obtained through an emic approach, is a first step for the development of a new generation of artificial cultural awareness supporting these systems.

Acknowledgement. We want to give a special thank to Eunika Laurent-Mercier, Associated Researcher in the Research Centre Magellan, University Jean Moulin, Lyon 3 for her support and advices.

References

1. E. G. Blanchard, R. Mizoguchi, and S. P. Lajoie, "Structuring the cultural domain with an upper ontology of culture," *The Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models*, pp. 179–212, 2010.
2. M. Rehm, Y. Nakano, E. André, T. Nishida, N. Bee, B. Endrass, M. Wissner, A. A. Lipi, and H.-H. Huang, "From observation to simulation: generating culture-specific behavior for interactive systems," *AI & society*, vol. 24, no. 3, pp. 267–280, 2009.
3. K. Reinecke and A. Bernstein, "Tell me where you've lived, and i'll tell you what you like: Adapting interfaces to cultural preferences," in *International Conference on User Modeling, Adaptation, and Personalization*, pp. 185–196, Springer, 2009.
4. A. Marcus and E. W. Gould, "Crosscurrents: cultural dimensions and global web user-interface design," *interactions*, vol. 7, no. 4, pp. 32–46, 2000.

[15] *http://www.legislation.gov.uk/ukpga/2010/15/contents*

5. P. Mohammed and P. Mohan, "Breakthroughs and challenges in culturally-aware technology enhanced learning," in *Proc. Workshop on Culturally-aware Technology Enhanced Learning in conjuction with EC-TEL 2013, Paphos, Cyprus, September 17*, 2013.
6. D. Maynard, A. Funk, and W. Peters, "Sprat: a tool for automatic semantic pattern-based ontology population," in *International conference for digital libraries and the semantic web, Trento, Italy*, 2009.
7. Z. Sellami, V. Camps, and N. Aussenac-Gilles, "Dynamo-mas: a multi-agent system for ontology evolution from text," *Journal on Data Semantics*, vol. 2, no. 2-3, pp. 145–161, 2013.
8. M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pp. 539–545, Association for Computational Linguistics, 1992.
9. W. H. Goodenough, "Report at the 7th annual round table meeting on linguistics and language study," 1957.
10. W. H. Goodenough, *Culture, language, and society*. Benjamin-Cummings Pub Co, 1981.
11. R. d'Andrade, "A folk model of the mind.," 1987.
12. W. A. Corsaro and D. R. Heise, "Event structure models from ethnographic data," *Sociological methodology*, pp. 1–57, 1990.
13. S. Stone-Jovicich, T. Lynam, A. Leitch, and N. Jones, "Using consensus analysis to assess mental models about water use and management in the crocodile river catchment, south africa," *Ecology and Society*, vol. 16, no. 1, 2011.
14. R. Mathevet, M. Etienne, T. Lynam, and C. Calvet, "Water management in the camargue biosphere reserve: insights from comparative mental models analysis," *Ecology and Society*, vol. 16, no. 1, 2011.
15. G. Bennardo and V. C. De Munck, *Cultural models: Genesis, methods, and experiences*. Oxford University Press, 2014.
16. C. Vuillot, N. Coron, F. Calatayud, C. Sirami, R. Mathevet, and A. Gibon, "Ways of farming and ways of thinking: do farmers' mental models of the landscape relate to their land management practices?," *Ecology and Society*, vol. 21, no. 1, pp. 1–23, 2016.
17. S. Wasserman and K. Faust, *Social network analysis: Methods and applications*, vol. 8. Cambridge university press, 1994.
18. M. Polanyi, *Personal knowledge: Towards a post-critical philosophy*. University of Chicago Press, 1958.
19. P. A. Alexander, D. L. Schallert, and V. C. Hare, "Coming to terms: How researchers in learning and literacy talk about knowledge," *Review of educational research*, vol. 61, no. 3, pp. 315–343, 1991.
20. N. Jones, H. Ross, T. Lynam, P. Perez, and A. Leitch, "Mental models: an interdisciplinary synthesis of theory and methods," *Ecology and Society*, vol. 16, no. 1, 2011.
21. C. S. Khoo and J.-C. Na, "Semantic relations in information science," 2006.
22. A. Wierzbicka, *English: Meaning and culture*. Oxford University Press, 2006.
23. W. Kempton, J. S. Boster, and J. A. Hartley, *Environmental values in American culture*. MIT Press, 1996.
24. Z. Oravecz, J. Vandekerckhove, and W. H. Batchelder, "Bayesian cultural consensus theory," *Field Methods*, vol. 26, no. 3, pp. 207–222, 2014.
25. W. H. Batchelder and A. K. Romney, "Test theory without an answer key," *Psychometrika*, vol. 53, no. 1, pp. 71–92, 1988.

26. A. K. Romney, S. C. Weller, and W. H. Batchelder, "Culture as consensus: A theory of culture and informant accuracy," *American anthropologist*, vol. 88, no. 2, pp. 313–338, 1986.
27. A. K. Romney, W. H. Batchelder, and S. C. Weller, "Recent applications of cultural consensus theory," *American Behavioral Scientist*, vol. 31, no. 2, pp. 163–177, 1987.
28. S. C. Weller, "Cultural consensus theory: Applications and frequently asked questions," *Field methods*, vol. 19, no. 4, pp. 339–368, 2007.
29. T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?," *International journal of human-computer studies*, vol. 43, no. 5-6, pp. 907–928, 1995.
30. W. N. Borst, *Construction of engineering ontologies for knowledge sharing and reuse*. Universiteit Twente, 1997.
31. A. Gòmez-Pérez and R. Benjamins, "Overview of knowledge sharing and reuse components: Ontologies and problem-solving methods," IJCAI and the Scandinavian AI Societies. CEUR Workshop Proceedings, 1999.
32. M. Fernàndez-Lòpez, A. Gòmez-Pérez, and N. Juristo, "Methontology: from ontological art towards ontological engineering," 1997.
33. R. Studer, V. R. Benjamins, and D. Fensel, "Knowledge engineering: principles and methods," *Data & knowledge engineering*, vol. 25, no. 1-2, pp. 161–197, 1998.
34. M. Pennacchiotti and P. Pantel, "Ontologizing semantic relations," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 793–800, Association for Computational Linguistics, 2006.
35. J. E. Mathieu, T. L. Rapp, M. T. Maynard, and P. M. Mangos, "Interactive effects of team and task shared mental models as related to air traffic controllers' collective efficacy and effectiveness," *Human Performance*, vol. 23, no. 1, pp. 22–40, 2009.
36. J. C. Young, "A model of illness treatment decisions in a tarascan town," *American Ethnologist*, vol. 7, no. 1, pp. 106–131, 1980.
37. K. Carley and M. Palmquist, "Extracting, representing, and analyzing mental models," *Social forces*, pp. 601–636, 1992.
38. N. Shuyo, "Language detection library for java," *Retrieved Jul*, vol. 7, p. 2016, 2010.
39. R. Girju, D. I. Moldovan, *et al.*, "Text mining for causal relations.," in *FLAIRS Conference*, pp. 360–364, 2002.
40. R. Girju, A. Badulescu, and D. Moldovan, "Learning semantic constraints for the automatic discovery of part-whole relations," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 1–8, Association for Computational Linguistics, 2003.
41. R. Girju, A. Badulescu, and D. Moldovan, "Automatic discovery of part-whole relations," *Computational Linguistics*, vol. 32, no. 1, pp. 83–135, 2006.
42. S. Caraballo, *Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text. Brown University Ph. D.* PhD thesis, Thesis, 2001.
43. S. Cederberg and D. Widdows, "Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy extraction," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 111–118, Association for Computational Linguistics, 2003.
44. P. Pantel and D. Ravichandran, "Automatically labeling semantic classes.," in *HLT-NAACL*, vol. 4, pp. 321–328, 2004.

45. T. Wang, Y. Li, K. Bontcheva, H. Cunningham, and J. Wang, "Automatic extraction of hierarchical relations from text," in *European Semantic Web Conference*, pp. 215–229, Springer, 2006.
46. G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
47. M. Hearst, "Wordnet: An electronic lexical database and some of its applications," 1998.

Appendix A Samples with their Respective Individuals

Sample	**Individuals**
Australian Police Forces	New South Wales Police Force, Northern Territory Police, Queensland Police, South Australia Police, Tasmania Police, Western Australia Police, Victoria Police
American State Police Forces	Arkansas State Police, Connecticut State Police, Delaware State Police, Idaho State Police, Illinois State Police, Indiana State Police, Kentucky State Police, Louisiana State Police, Maine State Police, Maryland State Police, Massachusetts State Police, Michigan State Police, New Hampshire State Police, New Jersey State Police, New Mexico State Police, New York State Police, Oregon State Police, Pennsylvania State Police, Rhode Island State Police, Vermont State Police, Virginia State Police
English Police Forces	Avon and Somerset Constabulary, Bedfordshire Police, Cleveland Police, Dorset Police, Essex Police, Greater Manchester Police, Hampshire Constabulary, Hertfordshire Constabulary, Lincolnshire Police, Nottinghamshire Police, Staffordshire Police, Suffolk Constabulary, Surrey Police, Sussex Police, Thames Valley Police, West Mercia Police, West Yorkshire Police, Wiltshire Police, Cambridgeshire Constabulary, Cheshire Constabulary, Cumbria Constabulary, Derbyshire Constabulary, City of London Police, Devon and Cornwall Police, Durham Constabulary, Gloucestershire Constabulary, Humberside Police, Kent Police, Lancashire Constabulary, Leicestershire Police, Merseyside Police, Metropolitan Police Service, Norfolk Constabulary, North Yorkshire Police, Northamptonshire Police, Northumbria Police, South Yorkshire Police, Warwickshire Police, West Midlands Police

Appendix B Individuals with their Number of Valid Sentences

Individual	**Number of Sentences**
New South Wales Police Force	66,362
Northern Territory Police	47,196
South Australia Police	28,303
Western Australia Police	58,872
Victoria Police	57,583
Connecticut State Police	81,755
Idaho State Police	50,163
Illinois State Police	111,931
Indiana State Police	37,341
Louisiana State Police	21,256
Massachusetts State Police	174,314
New Jersey State Police	135,880
Oregon State Police	40,339
Rhode Island State Police	29,214
Virginia State Police	79,426
Bedfordshire Police	42,942
Dorset Police	37,022
Essex Police	22,181
Greater Manchester Police	89,152
Hampshire Constabulary	41,300
Hertfordshire Constabulary	89,825
Lincolnshire Police	19,611
Nottinghamshire Police	75,610
Staffordshire Police	23,721
Thames Valley Police	65,234
West Mercia Police	45,964
Wiltshire Police	44,409
Cambridgeshire Constabulary	19,852
Cumbria Constabulary	16,572
Humberside Police	17,197
Lancashire Constabulary	12,504
Northumbria Police	60,931
South Yorkshire Police	35,128
West Midlands Police	10,499