Aligning Controlled Vocabularies using Faceted based approach

Ahsan-ul Morshed
Food and Agriculture Organization of the United Nations (FAO)
Rome,Italy
ahsan.morshed@fao.org

Margherita Sini
Food and Agriculture Organization of the United Nations (FAO)
Rome,Italy
margherita.sini@fao.org

Johannes Keizer
Food and Agriculture Organization of the United Nations (FAO)
Rome,Italy
Johannes.Keizer@fao.org

ABSTRACT
A vocabulary stores words, synonyms, word sense definitions (i.e. glosses), relations between word senses and concepts; such a vocabulary is generally referred to as the Controlled Vocabulary (CV) if choice or selections of terms are done by domain specialists. A facet is a distinct and dimensional feature of a concept or a term that allows taxonomy, ontology or controlled vocabulary to be viewed or ordered in multiple ways, rather than in a single way. The facet is also clearly defined, mutually exclusive, and composed by collectively exhaustive aspects, properties or characteristics of a domain. For example, a collection of rice might be represented using a name facet, place facet etc. In our case, we build a facet for each concept considering more general concepts (broader terms), less general concepts (narrow terms) or related concepts (related terms) that is to be called concept facet (CF). We use these CF's for mapping two controlled vocabularies. This methodology is based on hidden semantic matching which is different from the orthodox view of matching.

Keywords
Vocabulary Mapping,Thesaurus, AGROVOC, CABI

1. INTRODUCTION
The Semantic Web (which has gained widespread fame recently), where the underlying idea is that web contents should be expressed not only in natural language but also in a language that can be unambiguously understood, interpreted and used by software agents, thus permitting them to find, share and integrate information more easily. The central notation of the Semantic Web's idea is the ability to uniquely identify resources (with URIs) and languages (e.g. RDF/S, OWL) to formally represent knowledge (i.e. ontologies, which can simplistically be considered the taxonomies of classes representing objects, and of their inter-relationships) [13, 3]. These taxonomies contain domain knowledge; the domain is represented by a set of words and phrases used to describe concepts. A vocabulary is said to be controlled if it stores domain-specific chosen words, synonyms, word sense definitions (i.e. glosses) and relations between word senses and concepts [21]. In Controlled Vocabulary (CV), we denote the words as "blocks from which sentences are made", a synonym as "a word or phrase that refers to the same concept", a sense as "a meaning of a concept" and a concept as "an abstract idea inferred or derived from specific instances". The importance of CVs can hardly be underestimated; generally, each company or research group has its own information source e.g. databases, schemas and structures. Each of these sources has their respective set of individual CVs, creating a high level of heterogeneity. On one hand this is desirable, as it allows the involved parties to structure knowledge in a way which best fits their needs, e.g., for specific inter-office applications. On the other hand, individuals or companies also sometimes need a unified knowledge base (made up of different information sources) in order to satisfy their goals. This source of integration process requires a mapping between different CVs. Mapping between two CVs is generally a critical challenge for semantic interoperability. These CVs are used a lots as background knowledge for this data integration [8, 6]. What is more, classifications are matched using CVs are lightweight ontologies, also called Formal Classification (FC). In FC, lexical labels are translated to logical labels that remove ambiguities of natural language. For interested reading, we refer to [10, 7]. In our case, we are interested in the correspondence between concepts from two CVs, e.g., concept-to-concept mapping which includes word-to-word mapping, or synonym-to-synonym mapping. This mapping cannot be accomplished solely by a lexical comparison of two concepts using element level matcher [11, 14] that is included in SMOADistance, HammingDistance, JaroMeasure, SubStringDistance, N-gram, JaroWinKlerMeasure, and LavesteinDistance; we also need to consider the existing semantics. In light of the above discussion, the objective of this work is to determine a fully-automated mapping between two CVs and this work may be useful for navigating vocabularies, information extraction and linking information. Our this paper is an extension of our previous paper in details [1].

2. FACETED CONTROLLED VOCABULARY

2.1 Facet
A facet is like a diamond that is consisting of different faces. Its distinct features allow thesauri, classifications or tax-

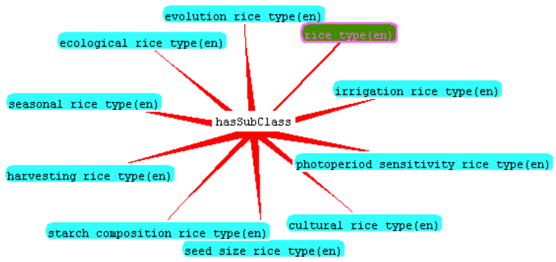

Figure 1: Rice Type

onomies to be organized in different ways, rather than in a single ways. The facet is also clearly defined, mutually exclusive, and composed by collectively exhaustive aspects of properties or characteristics of a domain. For example, a collection of rice might be classified using cultural and seasonal facets.

A Facet is constructed by following two steps [8]:

1. **Domain analysis:** First analysis of the term by consulting domain experts. This process is called idea plane, the language independent conceptual level, where simple concepts are identified. Each identified concept in expressed in the verbal plane is a given language. For example in English, trying to articulate the idea coextensively, namely identifying a term which exactly and unambiguously expresses the concept.

2. **Term collections and organization**: Secondly, collect terms and make an order of homogenous terms according to their characteristics, and order them (in hierarchies) in a meaningful sequence. The set of homogenous terms form a facet. For example, cow and milk form a facet called Dairy System (these entities are part-of relation with Dairy System).

Above steps construct a faceted knowledge organization system and corresponding to background knowledge, namely the a-priori knowledge which must exist in order to make semantic effective. Notice that the grouping of terms of step 2 have real world semantics, namely they are ontologies, classification and thesauri which are formed using partOf, isA, isSubClassOf and instanceOf relationships.

To properly consider a facet we need to consider the following point:

Specific characteristics of a domain topics can be seen as independent modularization of that domain. For instance, dairy product can be seen in Nutrition.

S.R. Ranganathan [15, 16] was the first to present the notion "facet" in library and information science (LIS). He proposed five different aspect to consider for building facet, PMEST: Personality (P), Matter (M), Energy (E), Space (S) and Time (T). However, his student Bhattarcharyya [2] proposed a refinement which consist of four main categories, called DEPA: Discipline (D) (what we call now a domain), Entity (E), Property (P), and Action (A).

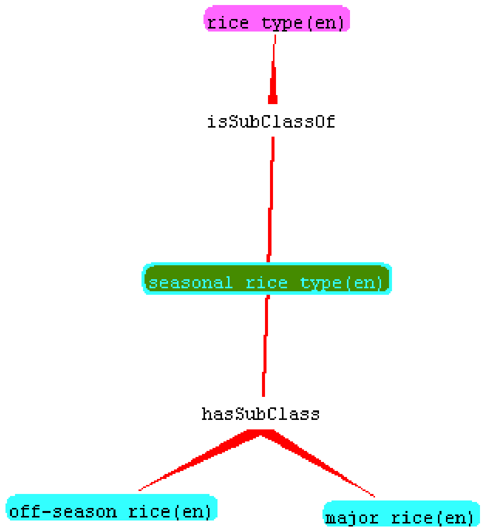

Figure 2: Seasonal Type Rice

In details DEPA can be picturised in the following way:

- Discipline (Domain): it includes established field of studies (e.g., Library Science, Mathematics and Physics), applications of traditional pure disciplines (e.g., Engineering and Agriculture), any aggregates of such fields (e.g., Physical Science and Social Sciences), or also more modern terms, fields like music, sports, computer science, and so on.

- Entity: the elementary category Entity is manifested in conceptual existence. Basically the concept represents the core idea of a domain treated as under this element category. For example: Rice is entity or concept in Agriculture domain.

- Property: it includes characteristic denoting quantities or qualitative characteristics. For example, quality, quantity, Measure, Weight, Taste, etc.

- Action: every concept should be considered with the notion of "doing". It includes processes and steps of doing. An action can manifest as "Self-action" or "External action" which is an action done by some agent (explicit or implicit) on or by itself. For example, Imagination, Interaction, Reaction, Reasoning, Thinking, etc. An external action is an action done by some agent (explicit or implicit) in a concept of any of the elementary categories described above. For example, Organization, Cooperation, Classification, Cataloging, Calculation, Design, etc.

To build a concept facet, we take discipline and then entity from DEPA model. Other properties will not be considered in this case. This process can be called semantic factoring. For example, we choose domain or discipline as Agriculture science. In this domain rice is entity or concept. Different kind of rices are existing in the world. Figure 1 [4] shows a distinct module of rice type which is lying in seasonal rice type, cultural rice type, seed size rice type and so on. These types depend on cultural, size, seasonal and others factors.

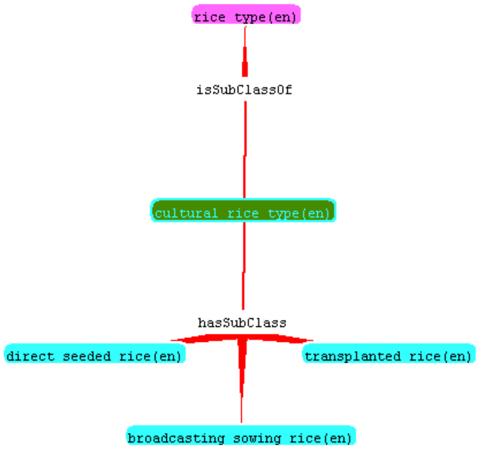

Figure 3: Cultural Type Rice

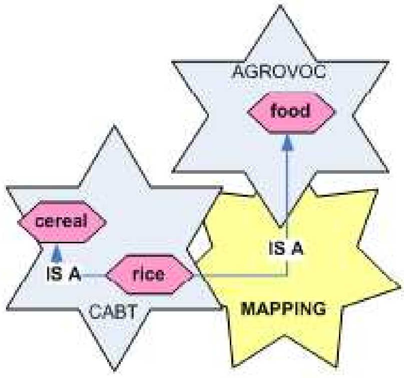

Figure 4: CV Matching

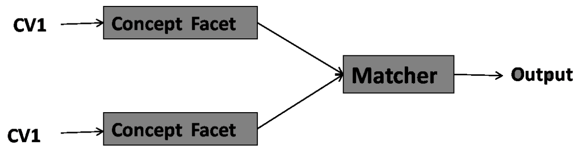

Figure 5: CV Matching

Each of which can be considered different facet.

Figure 2 shows one module of rice type which is seasonal type of rices. These type of rices are mostly cultivated in asian countries like India, Bangladesh, Nepal and Pakistan. These kind of rices are cultivated during rainy season and it comes out after two or three months. It is totally depended on time factor. Figure 3 shows Cultural type of rice. This class of rice mostly are cultivated in Thailand. Seeds are cultivated one time in the one place of land. After that it comes out from seeds directly; this kind of rice is called direct seed rice. On the other hand, some seeds are cultivated two times. One place is for growing a little part of seeds and then another place is for full growing and comes paddy; these kind of rice are called transplant rice.

3. CONTROLLED VOCABULARY MATCHING

Our problem revolves around the concept of CV matching base on the semantic matching idea described in [9]. The key intuition behind matching controlled vocabularies is the determination of mapping by computing syntactic and semantic relations which hold between the entities of any given two CVs [9, 19]. Let us consider matching 4-tuples $\langle ID_{i,j}, c_i, d_j, R \rangle$, $i = 1, ..., N_C$; $j = 1, ..., N_D$ where $ID_{i,j}$, is a unique identifier of the given mapped element; c_i is the i-th node of the CV1, N_C is number of nodes in the CV1, d_j is the j-th node of the CV2, N_D is the number of nodes in the CV2 and R specify a semantic relation which may hold between the concepts at nodes c_i and d_j. Therefore, light of the above discussion, the CV matching is defined with the following in problem: given two CV T_C and T_D compute the $N_C \times N_D$ mapped element $ID_{i,j}$, c_i, d_j, R with $c_i \in T_C$, $i = 1, ..., N_C$, $d_j \in T_D$, $= 1, ..., N_D$ and R is the strongest semantic relation holding between *concepts at node* c_i, d_j. Since we look for the $N_C \times N_D$ correspondence, the cardinality of mapping between elements can be determined to be $1 : N$. If necessary, these can also be decomposed straightforwardly into mapping elements with the 1:1 cardinality. For example:

We can find out the relationship between cereal and food if we have a mapped vocabularies.

4. CONCEPT FACET MATCHER

A Concept Facet (CF) contains distinct features for each concept: it includes combined relations, CF= $\langle lg, mg, R \rangle$, where lg identifies less general concepts (one or more), mg identifies more general concepts (one or more) and R identifies related concepts (one or more). In order to realize a matching between two vocabularies (CV1, CV2), we consider the CF from all given CVs's concepts: for every CF of CV1, we check the matching with all CFs of CV2. These concept facets are stored in tables for matching purpose. The methodology of the matching algorithm applied to every concept, can be represented with the following picture.

The matching between two concept facets follows the top-down approach and used several lexical comparison algorithms [11, 14] (SMOADistance, HammingDistance, JaroMeasure, SubStringDistance, N-gram, JaroWinKlerMeasure, and LavesteinDistance) . Firstly, we start comparing the more general concepts; if they match (they have same lexicalizations or they are synonyms) we assume that the concepts under investigation belongs to same concept (they match). Secondly (either we got match or not), we start comparing the less general concepts. Based on the results of two mentioned matching, we may obtain exact match (in case more general and less general concepts match), partial match (in case of only one match), or not match. Related concepts of CFs are considered to validate the previous results.

In short, we can express our CF matching algorithm in the following way:

In algorithm 1, we took each controlled vocabulary and stored each concept information in cF. cF is containing more

Algorithm 1 buildCFacet(CV)

```
for i = 0 to CV do
    store cF ← (Mg,Lg,R)
end for
return cF
```

general concepts (BT), less general concepts (NT) and related concepts (RT).

Algorithm 2 MatchingFacet(CV1,CV2)

```
cF1=BuildCFacet(CV1)
cF2=BuildCFacet(CV2)
for i = 0 to cF1 do
    for j = 0 to cF2 do
        cfmatcher=elementLevelMatcher(cF1,cF2)
    end for
end for
```

In algorithm 2, we compare two concept facets using element level matchers and store all matching information in cfmatcher.

5. RESULTS AND EVALUATION: THE AGROVOC AND CABI CASE STUDY

In our experiments, we used the AGROVOC thesaurus and the CABI thesaurus because there is no complete mapping between them. The results of the mapping will be published online so that users can use them for better indexing, searching and information retrieval [12, 20].

5.1 AGROVOC

AGROVOC is a multilingual controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains (e.g. the environment). The AGROVOC Thesaurus was developed by FAO and the Commission of the European Communities in the early 1980s. Since then it has been updated continuously by FAO and local institutions in member countries. It is mainly used for indexing and retrieval data in agricultural information systems both inside and outside FAO. It has approximately 20,000 concepts and four types of relations derived from the ISO standard. Among the available format, we used the XML version for our task [17].

5.2 CABI

CABI is a monolingual controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, horticulture, soil science, entomology, mycology, parasitology, veterinary medicine, nutrition and rural studies. The CABI thesaurus was developed by CABI which is a not-for-profit, science-based development and information organization. It has 48,000 concepts and four types of relationship derived from the ISO standard. We obtained data as text format and converted it to XML format for experiment purposes [18].

5.3 Results and Evaluation Descriptions

We started our experiments using 1000 concepts from each controlled vocabulary. Managing all concepts was a challenge because the two vocabularies are not organized in the same structure. We converted each vocabulary to the same format in order to conduct the test. We obtained 325 exact matches, 550 partial matches and 125 not matches concepts from FALCON-AO. Also, we obtained 175 exact matches from Concept Facet Matcher (CF-Matcher), but we found different numbers of partial matches from eight element label matchers. SMOADistance matcher gives more partial matches than others. Hamming distance, JaroMeasure, SubStringDistance, and N-gram do not give a satisfactory numbers of matches. JaroWinKlerMesaure and LevesteinDistance produce quite similar results. However, we got 465 partial matches (average) and 360 not matches (average) concepts from these element level matchers. In figure 6 we can see this results.

Further more, we chose FALCON-AO (Automatic Ontology Matching tool) because it had given the best results

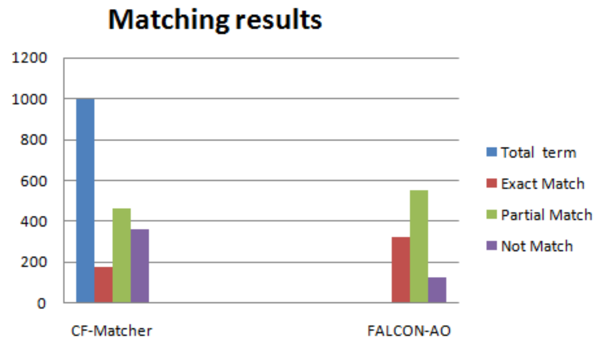

Figure 6: CV Matching Results

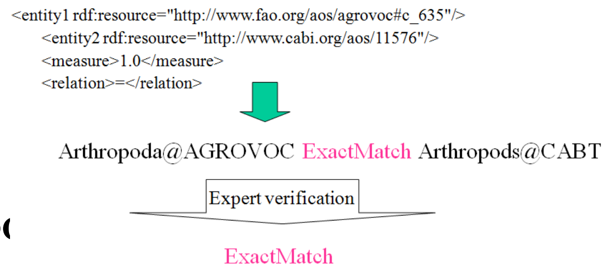

Figure 7: Exact Match

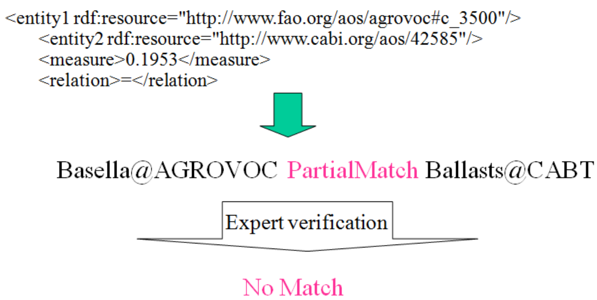

Figure 8: Partial Match

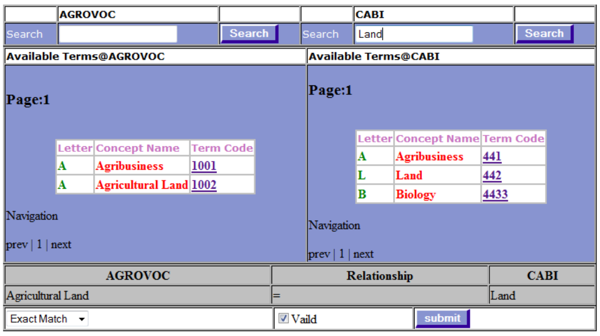

Figure 9: CV Matching System Prototype

according to mapping evaluation reports [5]. In our experiments, we considered 0.19 as our given threshold value for partial match and 1.0 for exact match. Figure 7 shows exact match between AGROVOC concept number c_635 and CABT concept number 11576. Similarly, figure 8 shows partial match between AGROVOC concept number c_3500 and CABT concept number 42585. We got these results from FALCON-AO and CF-Matcher. But according to our Domain expert at FAO, figure 7 shows correct results and figure 8 shows not correct results. Because there is no relationship between "Basella" and "Ballasts". From Dictionary, In Figure 9, we presented a human readable prototype so that one can access concepts information from two thesauri and see mapping results. The domain experts can validate the results and this information is stored into the database. We faced lots of challenges during our experiments. Overall, lots of data were overlapping and two automatic tools gave some partial matches which were not correct according to experts. However, our correct mapping results, after verification and validation, will be used for searching purpose

6. CONCLUSION

In this paper, we have shown our proposed system for automatic vocabulary matching using concept facets. We are convinced that it helps for better information searching, browsing, and extraction in agriculture and related domains. There are some open research issues: the semantic heterogeneity between two controlled vocabularies in a single domain; the multi-word concepts; the possibility of automatically link non-matched concepts to external reliable resources such as public thesauri, encyclopedia or dictionaries. Now, we are extending our work for semantic search for Agricultural domain.

Acknowledgment

Authors would like to thank Prof. Fausto Giunchiglia and Stephen Katz for their valuable suggestions. Also, we would like to thank Shaun Hobbs of CABI for kindly providing us with the data files.

7. REFERENCES

[1] A.Morshed and M.Sini. Creating and aligning controlled vocabularies. In *Advance Technology for Digital Libraries,AT4DL,Trento, Italy*, 2009.

[2] Bhattachary.G. Popsi:its foundamentals and procedure based on a general theory of subject indexing language. In *Libary Science with a slant to Documentation*, volume 16, pages 1–34, 1979.

[3] P. Bouquet, L. Serafini, and S. Zanobini. Semantic coordination: a new approach and an application. *In Proc. of the 2nd International Semantic Web Conference (ISWO'03). Sanibel Islands, Florida, USA*, October 2003.

[4] AGROVOC concept server. http://naist.cpe.ku.ac.th/agrovoc/.

[5] OAEI 2007 environment Trilateral Thesaurus Mapping Task. See http://www.few.vu.nl/ wrvhage/oaei2007/environment.html/.

[6] P. Shvaiko F. Giunchiglia and M. Yatskevich. Discovering missing background knowledge in onology matching. In *17th European Conference on Artificial Intelligance (ECAI 2006)*, volume 141, pages 382–386, 2006.

[7] M.Marchese F.Giunchiglia and I.Zaihrayeu. Encoding classifications into lightweight ontologies. *Data Semantics VIII*, pages 57–81, 2007.

[8] F. Giunchiglia, B.Dutta, and V.Maltese. Faceted lightweight ontologies. In *LNCS*, 2009.

[9] F. Giunchiglia, P. Shvaiko, and M. Yatskevich. S-match: An algorithm and an implementation of semantic matching. *In Proceedings of ESWS'04*, 2004.

[10] F. Giunchiglia and I. Zaihrayeu. Lighweight ontologies. *Technical report at DIT, the University of Trento, Italy*, October 2007.

[11] J.Euzenate and P.Shaviko. *Ontology Matching*. Springer, 1st edition, 2007.

[12] Sini M. Chang C. Li S. Lu W. He C. Liang, A. and J. Keizer. The mapping schema from chinese agricultural thesaurus to agrovoc. In Proceedings of the fifth Conference of the European Federation for Information Technology in Agriculture, Food and Environment and the thirdWorld Congress on Computers in Agriculture and Natural Resources, 2005.

[13] Bernardo Magnini, Luciano Serafini, and Manuela Speranza. Making explicit the semantics hidden in schema models. *In: Proceedings of the Workshop on Human Language Technology for the Semantic Web and Web Services, held at ISWC-2003, Sanibel Island, Florida*, October 2003.

[14] Natalya F. Noy. Semantic integration: a survey of ontology-based approaches. *SIGMOD Rec.*, 33(4):65–70, 2004.

[15] S.R Ranganathan. Element of library classification. Asia Publishing house.

[16] S.R Ranganathan. Prolegomena to library classification. Asia Publishing house.

[17] Agrovoc thesaurus. http://www.fao.org/agrovoc/.

[18] CAB thesaurus. http://www.cabi.org/.

[19] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[20] L.Finch H. Kolb W.Hage, M.Sini and G.Schreiber. The oaei food task:an analysis of a thesaurus alignment task.

[21] I. Zaihrayeu, L. Sun, F. Giunchiglia, W. Pan, Q. Ju, M. Chi, and X. Huang. From web directories to

ontologies: Natural language processing challenges. In *ISWC/ASWC*, 2007.