elpublishing

Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 9*. Editor: Julia Sallabank

Ontologies in language documentation

STEVE PEPPER

Ontologies in language documentation

Steve Pepper

1. Introduction

What is the role of ontologies in language documentation theory and practice? This paper clarifies the meaning of the term 'ontology' in the context of information management and the Web, and emphasizes the importance of distinguishing between knowledge representation and knowledge organization. It then examines how the term 'ontology' has been applied in the field of linguistics, focusing on a particular *kind* of ontology that is regarded as especially relevant in the context of language documentation. The General Ontology for Linguistic Description (GOLD) is presented in some detail, along with criticisms that have been raised against it. Finally it is suggested that the discipline of language documentation has more need for a knowledge organization system, and a shared thesaurus, than for an ontology-based knowledge representation system.

2. What is (an) ontology?

The term 'ontology' has become something of a buzzword during the last decade and been used to denote a number of quite different phenomena. The original usage is for what the *Shorter Oxford English Dictionary* calls 'the science or study of being; that department of metaphysics which relates to the being or essence of things, or to being in the abstract.'

The term was adopted by the Artificial Intelligence community in the 1980s to denote models of KNOWLEDGE REPRESENTATION used by intelligent agents, such as autonomous software programs designed to perform tasks that require human-like intelligence. One widely cited definition from this field is that of Gruber (1995), who states that 'an ontology is an explicit specification of a conceptualization'.[1]

[1] Gruber further defines conceptualization as 'the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them'.

Steve Pepper 2011. Ontologies in language documentation. In Julia Sallabank (ed.) *Language Documentation and Description*, *Vol 9,* 199-218 London: SOAS.

Another definition is given by Sowa (2000:492) in *Knowledge Representation*:

> The subject of ONTOLOGY is the study of the categories of things that exist or may exist in some domain. The product of such a study, called AN ONTOLOGY, is a catalog of the types of things that are assumed to exist in a domain of interest *D* from the perspective of a person who uses a language *L* for the purpose of talking about *D*. (emphasis added).

The publication of Tim Berners-Lee's vision of the 'Semantic Web' in 2001 popularized the idea of ontologies for a wider audience, and the subsequent development of Web Ontology Language (OWL) by the World Wide Web Consortium (W3C) has led many people to identify the concept of 'ontology' with OWL. However, OWL is just one of a number of formal languages that can be used to express an ontology; others include CycL, F-Logic, KIF and KL-ONE.

In a seminal contribution to the discussion of ontologies in the context of the Semantic Web, McGuinness (2003) acknowledges that a broad interpretation of Gruber's definition would include controlled vocabularies, glossaries and thesauri – that is, models of KNOWLEDGE ORGANIZATION traditionally associated with library science – but she chooses to exclude these from consideration. Her requirements for regarding something as a 'simple ontology' are that it should exhibit the following minimal set of properties:

- Finite controlled (extensible) vocabulary
- Unambiguous interpretation of classes and term relationships
- Strict hierarchical subclass relationships between classes

Garshol (2004) views ontologies in terms of subject-based classification and sees them as the highest stage in an expressivity continuum that starts with controlled vocabularies (a flat set of terms) and progresses through taxonomies (in which terms are arranged hierarchically) to thesauri (which add certain associative relationships[2]). The latter are all 'fixed-vocabulary languages'. Ontologies, by contrast, have open vocabularies that allow the creator of the subject description language to 'define the language at will'.

[2] Viz. RT, which expresses relationships between 'related terms' (cf. 'see also' in a back-of-book index), and USE and UF ('use for'), which express relationships between alternate terms and a preferred term (cf. 'see' in a back-of book index).

3. Ontologies in linguistics

Nickles et al (2007: 35) identify three fields of linguistics in which the term ontology is applied. The first of these – COMPUTATIONAL LINGUISTICS – uses ontologies in applications such as machine translation, information extraction, question answering, human-computer dialog systems, and text summarization, none of which are immediately relevant to language documentation. The second – MODEL-THEORETIC FORMAL SEMANTICS – is concerned with applying logic to find answers to the question 'What kinds of things do people talk as if there are?' This, too, would appear to have little direct relevance in language documentation.

It is the field dealing with LINGUISTIC TERMINOLOGY which is of most relevance to language documentation. This field has to keep pace with 'all the terminological innovations that keep growing in the different schools of linguistics around the globe' and therefore requires an 'ontology for linguistics' (Nickles et al 2007:39):

> The notion 'ontology for linguistics' refers to those conceptualizations of the domain of language and languages that are used to 'talk linguistics', to express and describe linguistic phenomena with the help of the corresponding concepts and the relations between them. The linguistic codings of these concepts are often, but by no means exclusively, technical terms of linguistics.

Two reasons are given why such an 'explicitly spelled-out' and 'well-defined' ontology is urgently required:

1. Precise descriptions of linguistic phenomena without precisely defined technical terms are impossible.
2. Only with the help of these tools can linguists reliably compare and compile different descriptions within a language and across languages.

As the authors point out, the first of these is a truism, since any precise and scientific description is by definition based on a set of clearly-defined terms and concepts. Thus (1) cannot be said to justify the need for a *shared* set of concepts, which is what is really at issue here. On the other hand, (2) clearly presupposes concept sharing, since it both involves 'different' descriptions (presumably by different linguists) and also has a cross-linguistic aspect. However, comparison and compilation are two quite different things, and whereas the latter is definitely relevant in language documentation, it could be argued that the former is mostly of interest to typologists.

4. General ontology for linguistic description

Nickles et al cite two projects that are working on 'ontologies for linguistics': the General Ontology for Linguistic Description (GOLD) and the Domain Ontology for Linguistic Phenomena (DOLPhen). The latter is based on a conception of language as a 'general purpose unbounded mind sharing device' (Zaefferer 2007:196), the consequence of which is the need for an 'ontology of everyday life' (Zaefferer 2007:215):

> Mind-sharing … presupposes shared systems of concepts, in other words, shared ontologies. Therefore an ontology has been proposed that is meant to include the most basic building blocks of everyday life conceptualizations which are reflected in everyday language.

Embedded within this General Ontology of Everyday Life, are two further ontologies: a Domain Ontology of Mental Entities, and DOLPHen itself, the latter strongly oriented toward a formal description of speech acts. It is thus more relevant to the field of pragmatics than language description, and indeed, from the available extracts the ontology does not seem to cover any of the fundamental concepts used in field of language documentation and description.

GOLD, on the other hand, is explicitly designed to be 'an ontology for descriptive linguistics'. It claims to provide 'a formalized account of the most basic categories and relations…used in the scientific description of human language'. Furthermore, it is 'intended to capture the knowledge of a well-trained linguist, and can thus be viewed as an attempt to codify the general knowledge of the field' (http://www.linguistics-ontology.org/info/about). That being the case, its relevance to language documentation is clearly worthy of consideration.

4.1. Background

GOLD emerged 'somewhat unexpectedly' (Simons & Hughes 2006) out of the E-MELD project, which had a two-fold objective:

1. To aid in the preservation of endangered languages data and documentation.
2. To aid in the development of the infrastructure necessary for effective collaboration among electronic archives.

GOLD was originally conceived as standard vocabulary for linguistic concepts that would solve the problem of disparate markup schemes for linguistic data (in particular data from endangered languages). However it was soon realized that such a one-size-fits-all solution would have no chance of adoption and GOLD was reconceived as a ‘conceptual ontology’ to which disparate vocabularies could be mapped. (Simons & Hughes 2006)

The current version of the ontology is GOLD 2008 and it can be browsed in HTML form at http://linguistics-ontology.org/gold/2008. It can also be downloaded as XML or OWL from http://linguistics-ontology.org/version.

4.2. Structure and content [3]

The ontology consists of approximately 500 classes arranged hierarchically within a top-level branching consisting of `Abstract`, `Object` and `Process`.

`Process` is as yet unpopulated and `Object` has only a small number of subclasses, shown in Figure 1 (together with the top-level classes of `Abstract`).

Figure 1: The complete GOLD `Object` hierarchy and the top-level of `Abstract`

```
Object                              Abstract [top-level only]
   LinguisticExpression                Character
      SignedLinguisticExpression       LinguisticDataStructure
      SpokenLinguisticExpression       LinguisticProperty
      WrittenLinguisticExpression      LinguisticSign
         OrthographicPart              LinguisticSystem
            Diacritic                  LinguisticUnit
            Digraph                    OrthographicSystem
            Glyph                      PhonologicalSystem
               Ligature                Taxon
         OrthographicPhrase
            OrthographicSentence
         OrthographicWord
          Paragraph
```

[3] This section describes the structure and content of GOLD. Portions of the ontology are reproduced *in extensio* in order to convey an impression of the amount of detail that it contains and make it possible to gauge its potential usefulness in language documentation.

Abstract is by far the most populous branch. Some of its almost 500 classes are nested up to 10 levels in depth. An extreme example of nesting is provided by the class Alveolar (Figure 2). It should be immediately apparent that the purpose of such a hierarchy is not to enable navigation by humans, but rather to provide a SUBSUMPTION HIERARCHY (what McGuinness terms a 'strict hierarchical subclass relationships between classes') of the type used by software agents to perform automated inferencing.

Figure 2: The hierarchical position of Alveolar

```
Abstract
   LinguisticProperty
      PhoneticProperty
         ArticulatoryProperty
            SupraLaryngealProperty
               PlaceProperty
                  BuccalProperty
                     CoronalProperty
                        AnteriorProperty
                           Alveolar
```

Less extreme in terms of nesting depth, but more impressive in terms of the sheer number of subclasses is CaseProperty. It boasts 55 subclasses, all of which are leaf nodes in the hierarchy (i.e. they are not themselves subdivided into further subclasses):[4]

Case (55): Abessive, Ablative, Absolutive, Accusative, Adessive, Allative, Aversive, Benefactive, Comitative, Contablative, Contallative, Conterminative, Contlative, Dative, Delative, Elative, Ergative, Essive, Genitive, Illative, Inablative, Inallative, Inessive, Instrumental, Interablative, Interallative, Interessive, Interlative, Interminative, Interterminative, Intertranslative, Intranslative, Lative, Locative, Malefactive, Nominative, Oblique, Partitive,

[4] Names of classes have been abbreviated by removing suffixes such as -Property, -Case, etc.

Perlative, Possessed, Subablative, Suballative, Subessive, Sublative, Subterminative, Subtranslative, Superablative, Superallative, Superessive, Superlative, Superterminative, Supertranslative, Terminative, Translative, Vocative

TAM properties are also well represented in the ontology:

Tense (32): CloseFuture, FutureInFuture, FutureInPast, FuturePerfect, Future, HesternalPast, HodiernalFuture, HodiernalPast, ImmediateFuture, ImmediatePast, NearFuture, NonFuture, NonPast, PastInPast, PastPerfect, Past, Perfect, PostHodiernalFuture, PreHodiernalPast, PresentPerfect, Present, RecentPast, Recent, RelativeFuture, RelativePast, RelativePresent, RemoteFuture, RemotePast, SimpleFuture, SimplePast, SimplePresent, Still

Aspect (17): Completive, Continuous, Distributive, Durative, Frequentive, Habitual, Imperfective, Inceptive, Iterative, NonProgressive, Perfective, Phasal, Progressive, Quantificational, Semelfactive, Simultaneous, Terminative

Mood (8): Dubitive, Indicative, Irrealis, Optative, Prohibitive, Realis, Subjunctive, Timitive

Other morphosyntactic categories, in addition to the above, are Voice (30), Modality (14), Number (12), Evidentiality (12), Force (10), Gender (9), Person (7), Polarity (2), Size (2) and Evaluative (2).

The non-morphosyntactic categories PartOfSpeechProperty, PhoneticProperty and HumanLanguageVariety further illustrate the breadth and depth of GOLD's coverage:

PartOfSpeech (61): Adverbial, Adverbializer, Auxiliary, CardinalNumeral, CoVerb, CommonNoun, ComparativeAdjective, Complementizer, ConVerb, Conjunction, Copula, Copulative, CorrelativeConnective, DefiniteArticle, Demonstrative, Disjunction, DistributiveNumeral, DitransitiveVerb, ExistentialMarker, Expletive, Gerund, IndefiniteArticle, IndefinitePronoun, Interjection, InterrogativeOperator, InterrogativeProform, IntransitiveVerb, Modal, MultiplicativeNumeral, NegationOperator, Nominal, NominalClassifier, NominalParticle, NounClassifier, NumeralClassifier, OrdinalNumeral, Participle, PartitiveNumeral, PersonalPronoun, PlainAdjective, PossessivePronoun, Postposition, Predicative, Prenoun, Preposition, Preverb, Proadjective, Proadverb, ProperNoun, Proverb, ReciprocalPronoun, ReflexivePronoun, RelativePronoun, Relativizer, Substantive, SuperlativeAdjective, SyntacticArgument, TransitiveVerb, VerbalAdjective, VerbalParticle, ZeroPlacePredicator

Phonetic (40): AcousticProperty, Alveolar, Apical, Approximant, Aspirated, Back, Breathy, Central, CentralEscape, Closed, Compressed, Creaky, Dental, Fricative, Front, GlottalMovementProperty, High, Laminal, LateralEscape, Low, Mid, MinusATR, MinusClick, MinusFortis, MinusNasal, ModalVoice, PlusATR, PlusClick, PlusFortis, PlusNasal, Postalveolar, Protruded, Retracted, Stop, Sublaminal, Tap, Trill, Unaspirated, Voiced, Voiceless

HumanLanguageVariety (10): AttestedVariety, DescribedVariety, ExtinctVariety, LivingVariety, NearlyExtinctVariety, SecondLanguageOnlyVariety, SignedLanguage, SpokenLanguage, UnattestedVariety, WrittenLanguage[5]

[5] An anonymous reviewer comments that this list could easily be criticized by an endangered languages expert.

Of the 518 classes in GOLD, 471 are annotated with the RDF Schema comment property. Approximately 70 of these are comments; the remainder are definitions, half of which contain references to sources. Table 1 in the Appendix lists the definitions for subclasses of `GenderProperty` which is fairly representative. It shows considerable inconsistency, both in terms of which concepts are documented, how they are documented, and the use of references.[6]

In addition to classes, GOLD 2008 defines a number of properties. There are seven datatype properties[7] (`abbreviation`, `hasExample`, `hasPageInformation`, `orthographicRep`, `phonemicRep`, `phoneticRep` and `stringRep`), of which only abbreviation is documented, so we assume these properties are still experimental.

More interesting are the 76 object properties which allow relationships to be expressed between individuals. Table 2 in the Appendix shows a representative selection and gives an idea of the kinds of relationship that the ontology seeks to sanction (as well as the current status of the documentation).

4.3. Purpose and use

GOLD was originally conceived as a 'morphosyntactic annotation inventory and label mapping scheme' before being formalized as an 'ontology by which disparate data sets can be integrated through a common representation of the basic linguistic features' (Simons & Hughes 2006). Then, in November 2004, the GOLD Community was formed with the following vision:

> By agreeing on a shared ONTOLOGY of linguistic concepts and on a shared infrastructure for INTEROPERATION, the linguistics community will be able to produce RESOURCES that describe individual languages in a comparable way, to develop TOOLS that produce these comparable resources, and to query SERVICES that aggregate as many comparable resources as are available (cited by Simons & Hughes 2006).

Three years later Farrar & Lewis (2007) proposed the GOLD Community of Practice (GOLDComm) as 'a model for linking on-line linguistic data to an

[6] The browsable version of the ontology also includes a number of examples, which are not part of the ontology.

[7] Datatype properties (renamed 'data properties' in the latest version of OWL) connect individuals with literals and thus correspond to attributes in other knowledge representation systems.

ontology', claiming that GOLD is 'a realization of the vision of the Semantic Web for descriptive linguistics'; the current About page of the GOLD website states that GOLD 'will facilitate automated reasoning over linguistic data and help establish the basic concepts through which intelligent search can be carried out.'

Farrar and Lewis, discussing the uses to which GOLD might be put, emphasize ontology-driven search as perhaps the most important application and distinguish two kinds of search: CONCEPT SEARCH and INTELLIGENT SEARCH. Searching by concept (rather than string value) improves both precision and recall: a search for `gold:PastTense` will not return documents concerning Pacific Standard Time, whereas a search for 'pst' will (increased precision); and a search for `gold:Subject` could return data that is marked for SUBJ, NOM and ERG as well as 'Subject' (increased recall).

Intelligent search goes one step further and infers meaning from the query:

> For example, if we pose the query 'List all the objects of verbs in Yaqui', the query engine could use the ontology to infer that by 'objects' we mean nouns (or noun phrases) since nouns are typically objects of verbs. It could also infer that nouns that are objects of verbs must be marked with a case appropriate to object position. In nominative/accusative languages like Yaqui, such a noun would be marked for accusative case. Thus, the search actually performed is 'List all instances of nouns marked for accusative case in Yaqui that are arguments of the verb'. (Farrar & Lewis 2007: 59)

One practical application which demonstrates some of the potential of GOLD is ODIN (http://odin.linguistlist.org/), the Online Database of Interlinear Text. ODIN is a database of interlinear glossed text (IGT) harvested from scholarly documents posted to the Web. It uses GOLD for term disambiguation, employing 'terminology sets' to map from the terminology used by the linguist to the corresponding concept in GOLD. (Lewis 2006)

4.4. Criticisms

A few (gentle) criticisms have been raised against GOLD, notably by Simons & Hughes (2006) and Munro & Nathan (2006), but the harshest comment is the lack of uptake in the community. Simons and Hughes put this down to 'three barriers':

- The complexity of the dissemination format which in effect places the threshold for engagement with GOLD at too high a level;

- The absence of a well defined change process through which GOLD can evolve into a standard that is truly community grounded;
- The lack of compelling GOLD-enabled applications which provide traction amongst end user communities.

According to Simons and Hughes, the expression of GOLD using OWL/RDF is an impediment because of its 'relatively complex representation' and they suggest using SKOS, the W3C's Simple Knowledge Organization System (an implementation of the standard thesaurus model), as the main distribution format.

Munro and Nathan are concerned that the ontology will be too inflexible and suggest that GOLD needs to explicitly support 'uncertainty, variability and phenomena that are inherently indeterminate or complex.' They also question GOLD's self-proclaimed theory-neutrality, pointing out that one of its primary assumptions (which explicitly rejects linguistic relativism) can hardly be called theory-neutral. Instead of theory-neutrality they argue for a '*pan*-theory model', one that allows variation. Their points are valid, but their proposed solution is open to question. They suggest representing each concept by a set of properties:

> A property would have three possible values to mark whether a given legacy ontology or language holds the property for a given concept: 'Yes', 'No', or 'Undefined' (default). For the ontology to accurately represent variance, it only needs to include enough properties to distinguish terms; however, for portability, it should seek to describe as many properties as possible.

There are two reasons for doubting whether this would work in practice. First of all, it would require an enormous amount of analysis to arrive at a suitable set of properties for defining each of the 500+ concepts in the existing ontology – far more than that required to write prose definitions. Secondly, and crucially, there is a recursion problem, since property types and values are also conceptual and would therefore require the same treatment, potentially ad infinitum.

A better solution, it seems, is to simply accept that every concept in the ontology is to some degree fuzzy. This would have no repercussions for the task of compilation. There *would* be repercussions for the task of comparison, if the intention is to use 'intelligent agents' rather than humans for this purpose, since such agents require clearly defined categories and subsumption hierarchies. But, as noted above, this is of minor interest in language documentation. The most serious repercussions, considered in the next section, are in terms of how we think about the approach being taken, the terminology we use to describe it, and the kinds of goals that can realistically be set.

5. Ontologies in language documentation

The preceding sections discussed the general nature of ontologies, and how they are currently used within linguistics, and one particular ontology was examined in some detail. In this section it is argued that the field of language documentation has very little practical use for ontologies in the 'true' sense of the term, and that the real need is for a set of commonly applied concepts that are easy to identify and easy to reference.

As Garshol (2004) makes clear, ontologies in the 'true' sense of the term transcend controlled vocabularies, taxonomies and thesauri. Those who originally borrowed the term from philosophy did so in order to denote something more than traditional models of knowledge *organization*. The key features that distinguish an ontology from, say, a thesaurus are its formality, its support for subsumption-based reasoning, and its extensible model (that is, the ability to define classes and relationships at will). These are the features that make it capable of knowledge *representation*, which involves capturing enough of the complexity of the real world for computers to be able to perform useful tasks on behalf of humans – over and above that of information collation and retrieval.

It is questionable whether language documentation has a need for such an ontology. The tasks involved in documenting languages – and the need to describe them in their own terms, unconstrained by a rigid, predefined set of concepts – require human judgement, and this precludes the use of inferencing engines. To the extent that computers can assist with language documentation tasks, it is in quite other ways. One is the automated production of interlinear glosses that is performed by tools like Shoebox; this requires a parser but not an ontology. Another is help in locating resources relevant to the language being documented (and in making such resources locatable by others); this does not require an ontology, but it does require the modern-day equivalent of a thesaurus.

Given the large number of undocumented languages, the increasing number of endangered languages, and the relatively meagre resources of the language documentation community, it is vital that existing resources be utilized as efficiently as possible. Time spent looking for information and duplicating the work of others should be minimized. This can be achieved by providing 'thick metadata' with language documentations (Nathan and Austin 2004). However, in order for this metadata to be effective, attention has to be paid to the kinds of values assigned to the metadata properties: as far as possible, those values need to be interoperable, and this indicates the use of a thesaurus.

A thesaurus is basically a set of terms and concepts used to specify the values of metadata properties such as 'keyword' (or its equivalent in the Dublin Core Metadata Initiative,[8] 'dc:subject'). Its purpose is simply to organize information by subject in order to make it easier for users to find what they are looking for. Terms are ordered hierarchically (and to some extent associatively, using the RT relation, see footnote 2), but the purpose of this is not to enable machine-based inferencing (which requires a subsumption hierarchy defined in terms of formal logic); the purpose is rather to provide a navigation aid for human users (i.e. those who are responsible for assigning metadata or who use it for searching).

Traditional thesauri have a number of limitations that need no longer apply in the modern world:

- The essentially monolingual approach of 'preferred terms' (USE and UF) is no longer appropriate in a still-emerging and globalized field in which the one-size-fits-all mentality cannot be applied to terminology. The ability of computers to use unique identifiers, leaving users free to choose their own terminology, solves this problem.[9]
- The emphasis on hierarchical navigation, originally dictated by maintenance considerations, does not reflect how people think (Bush 1945). In the age of digital hypertext, associative navigation, such as that offered by the Web, has become a much better solution.

Rather than an ontology (in the 'true' sense), what both language documentation and language typology need in order to support compilation is a set of common concepts for use as the values of thick metadata. Those concepts need to be inherently 'fuzzy' in order to address the concerns expressed by Munro and Nathan, and the set needs to be easily extensible on a user-defined basis (for obvious reasons). Each concept should be assigned a unique identifier to be used by computers when collating information about a common subject, leaving users free to choose their own terminology. And

[8] The DCMI (http://www.dublincore.org/) is one of the most widely applied metadata schemes.

[9] For example, elements carrying the GOLD identifier http://purl.org/linguistics/gold/Auxiliary can be variously labelled as 'Auxiliary verb', 'Hjelpeverb', 'Hilfsverb', 'Verb auxiliaire', etc. depending on the user's requirements.

those identifiers should be globally unique, in order to accommodate the needs of increasingly networked information.[10]

Hierarchical and associative structures can be superimposed on top of such concepts for the purpose of navigation and to improve findability, but the relations they express should *not* be construed as definitional in any sense (as they would be in an ontology), and users should be free to rearrange concepts into new hierarchies or create new associative navigation paths according to their needs.

Reassessing GOLD in the light of these kinds of requirements leads to the striking realization that this so-called 'ontology' is actually a good first approximation to the kind of thesaurus that is needed in the field of language documentation.[11] In particular, GOLD's coverage of the terminology of linguistic properties (used extensively in thick metadata) is very impressive. And for every one of these terms there is a globally unique identifier that could be added to a documentation at very little cost to the linguist.[12]

What is needed for GOLD to fulfill the role of a common knowledge organization system (KOS) for language documentation is the following:

1. Definitions need to be made fuzzier (or more prototypical) in order to enable linguists to talk in terms of 'roughly' the same concept, rather than 'precisely' the same concept.
2. The hierarchy needs to be relaxed and made more amenable to navigation by humans; the number of levels should be drastically reduced; parent-child relations should not be based on subsumption.
3. Additional concepts should be added to cover the most commonly used values for every kind of thick metadata property.[13]

[10] The published subjects paradigm (OASIS 2003, Pepper 2006) provides a model for creating and maintaining identifiers.

[11] It is thus not surprising that Simons and Hughes regard SKOS – a thesaurus model – as a more appropriate format than OWL.

[12] For example, language data exemplifying the use of auxiliary verbs (or the section of a grammar dealing with auxiliaries) could be assigned the metadata value GOLD:Auxiliary (an abbreviated form of http://purl.org/linguistics/gold/Auxiliary).

[13] Of course, the categories of thick metadata are essentially unconstrained and open ended, and therefore it must be possible for users to define identifiers for new categories as the need arises. But if information sharing is to take place, there must exist a widely accepted (and steadily expanding) set of identifiers that are used in common.

The last point requires some elaboration. Some examples of the kinds of concepts that are required in language documentation but not currently covered by GOLD are the following:

LANGUAGES. ISO 639-3 identifiers are not globally unique ('pst' can mean past tense and Pacific Standard Time, as well as Central Pashto). This problem can be solved by creating a concept in GOLD for each ISO 639-3 language, giving it an identifier consisting of the three-letter ISO code prefixed with a GOLD namespace (for Central Pashto, http://purl.org/linguistics/gold/iso639/pst), and providing a description that references ISO 639-3.

PEOPLE. Value would accrue in terms of discovery and reuse if people (e.g. collectors and speakers) were uniquely identified, rather than as, say, PKA, YM and LH (Nathan and Austin 2004:181). A KOS for language documentation could include identifiers for people, along with enough description to enable disambiguation.[14]

FORMATS AND ENCODINGS. Unique identifiers for concepts such as FOSF, XML, text file, Shoebox 5.0, Unicode, ASCII, UTF-8, etc. (examples from Nathan and Austin 2004) used as the values of metadata fields would improve search precision and recall.

OTHER METADATA VALUES. Any concept that is actually or potentially the value of a metadata field is worth including in a KOS for language documentation. Further examples inspired by Nathan and Austin (2004) are dictionary, finder list, monolingual, bilingual, trilingual, dictionary entry, headword, text collection, Open Access, etc.

TERMINOLOGY SPECIFIC TO LANGUAGE DOCUMENTATION THEORY, including the very concept 'language documentation' and the kinds of topics routinely discussed within the field: metadata (as a concept), particular metadata schemes (e.g. Dublin Core, OLAC), funding bodies, archives, initiatives and other organizations (e.g. HRELP, ELDP, ELAP, ELAR, DoBeS), publications (e.g. an identifier for the journal *Language*

[14] An anonymous reviewer comments: 'This seems futile, people can't be uniquely identified by names or any other natural attribute.' This is true, but it misses the point, which is that people *can* be uniquely identified by unique identifiers! For example, in the published subjects paradigm (see note 10), an identifier such as http://purl.org/linguistics/gold/people/Steven_Bird that is given in the text would resolve to a human-interpretable resource (say, a web page) that would provide enough information about the person in question to enable disambiguation in the event of several people bearing the same name. In this paradigm it is the duty of the identifier's *publisher* to ensure uniqueness, and the prerogative of the identifier's *user* to choose which publishers to trust when choosing among competing identifiers.

Documentation and Description that would distinguish it from the homonymous field of endeavour), various kinds of primary and secondary documentation (elicited sentences, stories, word lists, papers, grammars, dictionaries, etc.).

This list is far from exhaustive, but it should give an impression of the kind of concepts that could and should be included in a knowledge organization system for language documentation, and that would be considerably more useful than a formal ontology.

6. Conclusion

The term 'ontology' should be reserved for models that go beyond those of traditional knowledge organization and that enable the kind of knowledge representation required by intelligent agents to perform inferencing. This is not something that is of immediate use in language documentation.[15] Instead what is needed is a kind of thesaurus – a knowledge organization system – consisting of a set of concepts with globally unique identifiers that can be used as the values of thick metadata. In order to account for gradience, those concepts should not be defined more precisely than necessary, and any hierarchies into which they are organized should not be based on subsumption. Such a thesaurus would improve the findability of documentations and lead to more efficient use of resources. It would not necessarily improve their *documentation value* as such (except, possibly, in encouraging greater consistency), but it can be claimed that the value of a documentation – like that of any information – resides as much in its findability as in its actual content: a language documentation, whatever its quality, is of no value at all if its content cannot be located.

References

Berners-Lee, Tim, James Hendler & Ora Lassila. 2001. The Semantic Web. *Scientific American Magazine* http://www.scientificamerican.com/article. cfm?id=the-semantic-web&print=true (2010-11-26).

Bush, Vannevar. 1945. As We May Think. *Atlantic Monthly*, July 1945. http:// www. theatlantic.com/doc/194507/bush (2010-11-26).

[15] Nor is it of any value in language typology, unless one holds the position that languages can be fully described in terms of a precisely defined set of fixed (and thus presumably innate) concepts.

Farrar, Scott & D. Terence Langendoen. 2003. A linguistic ontology for the Semantic Web. *GLOT International* 7 (3), 97-100.

Farrar, Scott & William D. Lewis. 2007. The GOLD Community of Practice. An Infrastructure for Linguistic Data on the Web. *Language Resources and Evaluation* 41 (1), 45–60.

Garshol, Lars Marius. 2004. Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all. *Journal of Information Science* 30 (4), 378–391.

Gruber, T. R. 1995. Toward principles for the design of ontologies used for knowledge sharing. *International Journal Human-Computer Studies* 43 (5-6).

Lewis, W. D. 2006. ODIN: A model for adapting and enriching legacy infrastructure. In *Proceedings of the e-Humanities Workshop*, held in cooperation with *e-Science 2006: 2nd IEEE International Conference on e-Science and Grid Computing*, Amsterdam. http://faculty.washington. edu/wlewis2/papers/ODIN-eH06.pdf (2010-11-26).

McGuinness, Deborah L. 2003. Ontologies come of age. In Dieter Fensel, Jim Hendler, Henry Lieberman, & Wolfgang Wahlster (eds.), *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. Cambridge, MA: MIT Press. http://www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-mit-press-(with-citation).htm (2010-11-26).

Munro, Robert & David Nathan. 2005. Towards portability and interoperability for linguistic annotation and language-specific ontologies. *Proceedings of the E-MELD Workshop on Linguistic Ontologies and Data Categories for Language Resources (E-MELD 2005)* Boston. (2010-11-26)
http://www.robertmunro.com/research/munro05interoperability.pdf.

Nathan, David & Peter K. Austin. 2004. Reconceiving metadata: language documentation through thick and thin. *Language Documentation and Description* 2, 179–187.

Nickles, Matthias, Adam Pease, Andrea C. Schalley, & Dietmar Zaefferer. 2007. Ontologies across disciplines. In Schalley & Zaefferer (2007). http://uk.cbs.dk/content/download/95540/1243885/file/NicklesEtAl2007. pdf (2010-11-26).

OASIS. 2003. *Published subjects: Introduction and basic requirements.* http://www.oasis-open.org/committees/download.php/3050/pubsubj-pt1-1.02-cs.pdf (2010-11-26).

Pepper, Steve. 2006. *The case for published subjects*. Ontopia. http://www. ontopia.net/topicmaps/materials/The_Case_for_Published_Subjects.pdf (2010-11-26).

Pepper, Steve. 2010. Topic maps. In Marcia J. Bates and Mary N. Maack *Encyclopedia of Library and Information Sciences*, Third Edition. Boca Raton, FL: CRC Press.

Schalley, Andrea C. & Dietmar Zaefferer. 2007. *Ontolinguistics: How ontological status shapes the linguistic coding of concepts*. Berlin: Mouton de Gruyter.

Simons, Gary & Baden Hughes. 2006. GOLD as a standard for linguistic data interoperation: A road map for development. In *Proceedings of the EMELD'06 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art*. Lansing, MI. June 20-22, 2006. http://emeld.org/workshop/2006/papers/SimonsHughes.doc (2010-11-26).

Sowa, John F. 2000. *Knowledge representation: Logical, philosophical, and computational foundations*. Pacific Grove, CA: Brooks Cole.

Zaefferer, Dietmar. 2007. Language as mind sharing device: Mental and linguistic concepts in a general ontology of everyday life. In Schalley & Zaefferer (2007).

Appendix

Table 1. Definitions for subclasses of GENDERPROPERTY

AnimateGender: A grammatical gender property assigned to a class of nouns with animate denotation. In a given language it may include larger or smaller numbers of nouns which do not meet this semantic criterion. The animate gender may occur in a two-gender system, with the other gender being labelled inanimate. However, the animate gender may also occur in larger inventories (i.e. greater than two values). Examples of these larger systems are found in Bantu languages (where nouns denoting humans are included in the animate gender) and in languages of Daghestan (where the animate gender is typically for non-human animates) [Corbett 1991, 20-32].@en[16]

ArabicNumeralGender: [Need comment]

FeminineGender: A gender property established on the basis of agreement, to which nouns may be assigned if 1) they inherently denote females. Additionally, but not necessarily, nouns may be assigned this value if: 2) their formal properties (morphological or phonological) lead them to be assigned to the same agreement pattern as other nouns within the language that have female denotation. 3) they are arbitrarily assigned to the same agreement pattern as other nouns in the language that have female denotation [Corbett 1991].@en

HumanGender: [Need comment]

[16] @en appears to be a stray language tag that should not appear in the data.

InanimateGender: A grammatical gender property such that membership in the inanimate grammatical class is largely based on meaning, in that non-living things, such as objects of manufacture and natural 'non-living' things are included in it. For example, one of the two grammatical genders, or noun classes, of Nishnaabemwin, the other being animate [Valentine 2001, 114].
MasculineGender: *null*
NeuterGender: NeuterGender[17]
NeuterGender: A gender property established on the basis of agreement, to which nouns may be assigned, either by a semantic rule, if they belong to the semantic residue of the assignment system, or by a formal rule, if assignment depends on inflectional class membership. Typically, this means that the neuter gender may cover some inanimates and possibly some portion of lower order animates. Note: Although in familiar Indo-European languages the term neuter gender may be part of a system with three or less values, it can be used for systems containing more than three gender values (e.g. Bininj Gunwok).
RomanNumeralGender: *null*
VegetableGender: Vegetable gender refers to inanimates and exists in some four-way gender systems, e.g., masculine, feminine, neuter, and vegetable as in Bininj Gun-wok [Evans 2003, 202].

Table 2. Definitions for instances of OBJECTPROPERTY

acousticRealization: The relation between some linguistic unit and its corresponding spoken expression.
agrees: A relation holding between syntactic units, often manifesting itself in shared form features. NOTE: this could be better defined once syntactic roles and relations are developed.
allomorph: The relation that holds between a morpheme and one of its morphs, an occurrence of a morpheme in context.
allophone: The relation that holds between a phoneme and one of its phones, an occurrence of a phoneme in context.

[17] This duplicate comment is an error.

ancestorVariety. ancestorVariety is the predicate expressing the basic diachronic relationship between a language variety that existed some time in the past and a variety existing at a later time such that the former has evolved into the latter through regular language change.

coda: The closing segment of a syllable.

directObject: A direct object is a grammatical relation that exhibits a combination of certain independent syntactic properties, such as the following: the usual grammatical characteristics of the patient of typically transitive verbs; particular case marking; a particular clause position; the conditioning of an agreement affix on the verb; the capability of becoming the clause subject in passivization; the capability of reflexivization. The identification of the direct object relation may be further confirmed by finding significant overlap with similar direct object relations previously established in other languages. This may be done by analyzing correspondence between translation equivalents [Crystal 1985, 94; Hartmann and Stork 1972, 155; Comrie 1989, 66; Andrews 1985, 68,120,126; Comrie 1985, 337].

entailedBy: *null*

entails: *null*

feature: The relation between a linguistic unit and a linguistic feature. A feature inheres in its host. NOTE: this relation is distinct from the `hasFeature` which pertains to data structures.

follows: This relation holds between two linguistic units and represents the inverse of 'precedes'. That is, (follows A B) means that A comes after B in the linearization of the realization of linguistic signs. The inverse of this relation is 'precedes'.

freeTranslation: The relation between an orthographic expression in one language and some orthographic expression in another such that both expressions have exactly the same meaning. The words in the translation may not correspond to the those in the source expression.

geneticallyRelated: geneticallyRelated is the basic kinship relation between languages varieties. If two language varieties are genetically related, then this implies that both varieties are derived from a common proto-language.