Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5668–5677
Marseille, 11–16 May 2020

Annotation Interoperability for the Post-ISOCat Era

Christian Chiarcos, Christian Fäth, Frank Abromeit
Applied Computational Linguistics Lab (ACoLi)
Goethe University Frankfurt, Germany
{chiarcos, faeth, abromeit}@em.uni-frankfurt.de

Abstract

With this paper, we provide an overview over ISOCat successor solutions and annotation standardization efforts since 2010, and we describe the low-cost harmonization of post-ISOCat vocabularies by means of modular, linked ontologies: The CLARIN Concept Registry, LexInfo, Universal Parts of Speech, Universal Dependencies and UniMorph are linked with the Ontologies of Linguistic Annotation and through it with ISOCat, the GOLD ontology, the Typological Database Systems ontology and a large number of annotation schemes.

Keywords: interoperability, ontologies, linked data, linguistic annotation

1. Background

The divergence and heterogeneity of linguistic annotations even for comparable language resources for the same language variety has long been recognized as a key problem in the advancement of human language technology. The creation of language resources and tools is a laboursome and cost-intense process, but if these cannot be easily combined with each other, be it to feed the output of one annotator (say, a POS tagger) as input to another (say, a parser), or to increase the amount of available training data by combining multiple corpora of the same kind (say, morphology, part of speech annotations, dependency syntax or phrase structure syntax), potential synergies between those annotation efforts cannot be exploited and progress in the field is thus unnecessarily delayed. Very often, this is the situation for low-resource languages, where no standard annotation for a particular phenomenon has been established yet, but different researchers develop multiple schemas independently from each other.

With ISOCat,[1] the ISO TC37 Data Category Registry (DCR), a possible solution to this problem has been proposed in the 2000s, a central repository of linguistic terminology managed in the context of a community process with a relatively low entry barrier, where individual researchers or institutions would just register their resource- or application-specific terminology. ISOCat focused on the elementary level of linguistic terms, and deliberately excluded the relations between these terms, but eventually, it was imagined that these relations could be added in a separate relation registry (Schuurman and Windhouwer, 2011), and the process to develop these relations and terms would converge towards a consistent terminological inventory whose content would be the basis for a subsequent standardization process within ISO TC 37.

In the standardization approach, terminology harmonization is achieved by aggregation and consolidation across all possible user communities within a centralized, monolithic repository. An alternative approach emerging simultaneously to ISOCat was the idea of distributed terminology harmonization by creating links among independent, self-contained domain terminologies and between them and one or multiple 'upper models', especially by means of ontologies and Semantic Web technologies:[2] Different user communities formalize and provide their respective terminology in a stand-alone, self-contained ontology, and these ontologies are subsequently linked with each other by means of designated relations (e.g., rdfs:subClassOf, skos:broader, owl:equivalentClass, or owl:sameAs) between identical or near-identical concepts.

The general idea is probably best described in the title of a seminal paper by Dimitriadis et al. (2009): "How to integrate databases without starting a typology war". The fundamental insight is that so far, any approach to develop or to enforce standard terminology in linguistics was rejected by the majority of the scientific community, and given the degree of specialization in various branches of linguistics, this is unlikely to change.

The General Ontology of Linguistic Description (Farrar and Langendoen, 2010, GOLD), currently hosted at LinguistList, is a similar effort to formalize reference terminology as an ontology for the field of language documentation and descriptive linguistics. The authors of GOLD see its role as providing a lingua franca as a basis for annotation projects to map their data categories to in order to foster conceptual interoperability.

In natural language processing and corpus linguistics, a similar concept was implemented with the Ontologies of Linguistic Annotation (Chiarcos and Sukhareva, 2015, OLiA). Historically, OLiA originated as an OWL formalization of the EAGLES recommendations (Leech and Wilson, 1996), extended by a linking with GOLD and the morphosyntactic and syntactic profiles of ISOCat as well as definitions and introduced by various annotation schemes it was applied to. OLiA owes its continued relevance to its application beyond its original use case: It has been conceived as a meta-vocabulary for tagset documentation and cross-resource corpus querying, but with the emergence of the Linguistic Linked Open Data cloud since 2010 (Chiarcos et al., 2012a),[3] it evolved to become the primary vocab-

[1] http://www.ISOCat.org/

[2] In fact, an ontological formalization, or at least the use of Semantic Web technology had been a design concept in early days of the ISO Data Category Registry (Ide and Romary, 2004), but not adopted for the effective implementation of ISOCat.

[3] http://linguistic-lod.org

ulary to formalize linguistic annotations for Semantic Web applications and language resources in the web of data.

A fundamental insight of the 2000s was that the development and usability of widely used, shared annotation terminology for linguistic annotations must be based on web technologies, and in particular, resolvable URIs. The technical standard in this regard still remains to be ISOCat, which provides persistent URIs, even though these are redirected to a static dump now rather than the underlying live system. Since Chiarcos et al. (2012b), linked data has been recognized as a key element to facilitate language resource interoperability, and this trend intensified with the more recent trend to shift from XML tecnologies of JSON(-LD) in the language resource community. Accordingly, RDF dumps of resources developed on the basis of proprietary formalisms have been made available at an increasing rate.

Along with the improvement in structural (format and access) interoperability, also the content of all major terminology resources became increasingly harmonized. OLiA already provideded an indirect linking between ISOCat and GOLD, but a direct bridge between both resources was established and GOLD and ISOCat began to converge when the 2010 edition of GOLD was mirrored within ISOCat (Kemps-Snijders, 2010). However, this process, as well as the addition of a large number of tagsets and domain vocabularies, contributed to the emergence of terminological (near-)doublets. Without relational data structures to express identity or near-identity, or an effective community process to eliminate such doublets, the ISOCat repository came to house an increasing number of duplicate and near-duplicate records where even elementary concepts existed multiple times, distinguished by their respective owners and the wording of their definitions, but not by their label (e.g., 'part of speech' as DC-1345, DC-3747 and DC-5294; 'verb' as DC-1424 and DC-4949; 'dative case' as DC-1265 and DC-3148; 'past' as DC-1347 and DC-4966; 'masculine gender' as DC-1883 and DC-3312).

Among the linguistic terminology resources around 2010, ISOCat excelled as being the most widely used, most fundamental, and richest. Unfortunately, its popularity and continued growth ultimately led to its abandonment, as the unrestricted addition of new records, together with very weak facilities for expressing the relations between entries, produced a largely unstructured collection of redundant entries and near-synonyms.[4] In the original conception of ISOCat, a community process was envisioned to consolidate duplicate entries and to formulate consensus definitions. For the specific case of ISOCat, however, this process, however, never produced any concrete results as it remained hard to motivate researchers to engage in abstract work such as consolidating terms and definitions at this scale. External extensions of ISOCat, however, involved the development of full-fledged domain ontologies that were grounded in ISOCat profiles. This includes an ontology for lexical data structures (LexInfo, see below), an ontology for linguistic annotations (Chiarcos, 2010) and an ontology of language resource metadata (Zinn et al., 2012). Neither of these ontologies, however, were adopted by ISOCat administrators nor considered as a possible input to the ISOCat community process or its future development. To some extent, this was due to the technological choices made in ISOCat architecture which was designed on the basis of an application-specific, XML-based format, well en par with the state of the art in early 2000's language resource technology, but orthogonal to RDF technology.

But despite its influence in numerous branches of research, ISOCat failed in general to deliver on its promises and was eventually discontinued in 2014 (Schuurman et al., 2015).

2. Linguistic Annotation Terminology since 2010

ISOCat content remains available as a static resource only `http://ISOCat.tbxinfo.net/`, and two direct successor systems are being developed – along with other more recent efforts to harmonize linguistic data categories. Whereas ISOCat was relatively widely used, these efforts target more specific communities and use cases. These include terminologies that directly build on ISOCat, esp., the CLARIN Concept Registry (Sect. 2.1.), developed as a central component of the CLARIN infrastructure, and LexInfo (Sect. 2.2.), a vocabulary for the terminology of lexical-conceptual resources. We distinguish these terminology repositories from other recent standardization efforts (Sect. 2.3.) developed independently from ISOCat by communities that aim to create cross-linguistically compatible annotations such as syntax, morphosyntax and inflectional morphology.

2.1. CLARIN Concept Registry and DatCatWeb

As a replacement for ISOCat, Schuurman et al. (2015) introduced the CLARIN Concept Registry (CCR),[5] which is based on OpenSKOS (Brugman and Lindeman, 2012). They aimed to avoid the issues with ISOCat by allowing only CLARIN National Content coordinators to update the registry, and by requiring a "good definition" of a concept that is unique, meaningful, reusable and concise. However, progress on this effort is slow, and even at the time of writing this, basic concepts such as 'part-of-speech' have not reached the 'approved' status.

Simultaneously, ISO TC37 has been developing DatCatInfo as an ISOCat successor registry initially populated with ISOCat concepts (Warburton and Wright, 2020).[6] DatCatInfo is developed in close connection with the TermBase eXchange format (TBX). As of early 2019, 2,977 data categories (approximately half the DCs from ISOCat) have been migrated to DatCatInfo, and are currently undergoing continued revision in order to eliminate duplicates and establish a coherent view on the terminology.

The future division of labour between the CCR and DatCatInfo is not clear, although they clearly diverge and we may anticipate a specialization of the CCR for applications in language technology and a specialization of DatCatInfo

[4] ISOCat provided an inventory of hierarchical relations (dcif:isA), but these were optional and not enforced. Out of 682 data categories in the morphosyntactic profile, only one third (274) used hierarchical relations, whereas the majority was provided as an unstructured list.

[5] `https://www.clarin.eu/ccr`

[6] `http://www.datcatinfo.net`

for lexical and terminological resources. In any case, both systems will provide resolvable URIs (at the time of writing, only CCR does), and for domain-specific vocabularies, it will be possible to link them to each of them. In fact, the capability to facilitate linking with multiple external reference models has been the motivation for modular architectures such as OLiA (see below).

2.2. LexInfo

LexInfo[7] is the representative vocabulary for linguistic data categories for lexical-conceptual resources in the context of Linguistic Linked Open Data, especially because of its intrinsic ties with the popular OntoLex-Lemon vocabulary (Cimiano et al., 2016). Originally, it was designed as an ontology for "associat[ing] linguistic information with respect to any level of linguistic description and expressivity to elements in an ontology" (Cimiano et al., 2011). In this function, it predates Ontolex-Lemon, but with LexInfo v.2.0, it was re-designed to serve as a terminology backend of OntoLex-Lemon with the goal of making Ontolex-lemon itself agnostic of any linguistic category system. The LexInfo ontology developed out of an RDF edition of the Lexical Markup Framework (Francopoulo et al., 2006, LMF), i.e., a major source of ISOCat concepts, so that LexInfo is largely compatible with ISOCat. LexInfo provides an axiomatized set of linguistic categories, covering areas such as part of speech, tense, number, animacy, degree, mood, register, etc. These categories are largely derived from ISOCat, but LexInfo provides a stronger axiomatization and a coherent global organization.

LexInfo v.2.0 is the reference vocabulary for linguistic categories and features in lexical-conceptual resources in the web of data. Since December 2019, version 3.0 is in preparation,[8] with the goal to increase its cross-linguistic applicability and its compability with OntoLex-Lemon, a novel aspect here is that this development is conducted in the style of an open source project in order to facilitate the participation of the wider community.

2.3. Other Recent Standardization Efforts

The Universal Dependencies (Nivre et al., 2016, UD)[9] aim to provide cross-linguistically applicable annotations for dependency syntax, parts of speech and morphosyntactic features. UD differs from earlier standardization efforts in that it is coupled with the creation and the release of open source corpora with the corresponding annotations.

The Universal Dependencies have been embraced enthusiastically by the NLP and language resource communities due to coincidence with an increased interest in the syntactic annotation of low-resource languages, and so far, more than 100 treebanks in over 70 languages are being provided. Inspired by the wide success of UD, similar efforts have been undertaken for other areas of application, e.g., with UniMorph (Sylak-Glassman et al.,)[10] for morphology.

With ISOCat development stalled, and new standardization initiatives emerging, we see a great risk in increasing fractionalization in language resource development. With this paper, we provide a linking for the existing terminology repositories for linguistic data categories. We do not address language resource metadata, for which we refer the interested reader to a parallel effort, the development of METASHARE-OWL (McCrae et al., 2015), initiated at the 1st Summer Datathon on LLOD (SD-LLOD 2015), and currently continued in the context of the European Language Grid (ELG) and the Linked Data for Language Technology (LD4LT) Community Group of the W3C.[11]

3. Approach

With the goal of linking multiple, and increasingly fractionalizing vocabularies for linguistic data categories, we focus both on vocabularies grounded in ISOCat, i.e., LexInfo, CCR and OLiA, as well as on novel vocabularies for morphosyntactic and syntactic annotation developed independently, i.e., the Universal Dependencies vocabularies (Universal Parts of Speech, UD v.1, UD v.2) and UniMorph. Instead of mapping or linking each of them with every other, we use the OLiA Reference Model as an intermediate layer and thereby link them with older vocabularies that OLiA is grounded in, i.e., GOLD and ISOCat. We further describe how this linking can be used to derive mapping tables that can be used to facilitate interoperability in NLP applications and transformation tasks.

3.1. Ontologies of Linguistic Annotation

The Ontologies of Linguistic Annotation (Chiarcos, 2008; Chiarcos, 2010, OLiA) have been designed as a mediator between various terminology repositories on the one hand and linguistically annotated resources (more precisely, their annotation schemes), on the other hand (Schmidt et al., 2006). OLiA applies linked data principles to leverage several distributed terminology repositories: It provides the formalization of the mapping from annotations via the OLiA Reference Model to several existing terminology repositories ('External Reference Models') by means of a modular architecture of interdependent OWL2/DL ontologies: Annotation models and (external) reference models each constitute self-contained, standalone ontologies, whereas the linking between them is a physically separated ontology that imports the respective models and asserts rdfs:subClassOf (rdfs:subPropertyOf) relations about their concepts (and properties).

The OLiA ontologies are available from `http://purl.org/olia` under a Creative Commons Attribution license (CC-BY), and they are developed as an open source project using GitHub.[12]

Four different types of ontologies are distinguished (Fig. 1): (1) The OLiA Reference Model is an OWL ontology that specifies the common terminology that different annotation schemes can refer to. (2) Multiple OLiA Annotation Models formalize annotation schemes and tagsets. Fig. 1 illustrates this with an annotation model developed as part of the Korean NLP2RDF stack (Hahm et al., 2012). (3) For every

[7] `https://www.lexinfo.net/`
[8] `https://github.com/ontolex/lexinfo`
[9] `https://universaldependencies.org/`
[10] `http://unimorph.github.io/`
[11] `https://github.com/ld4lt/metashare`
[12] `https://github.com/acoli-repo/olia`

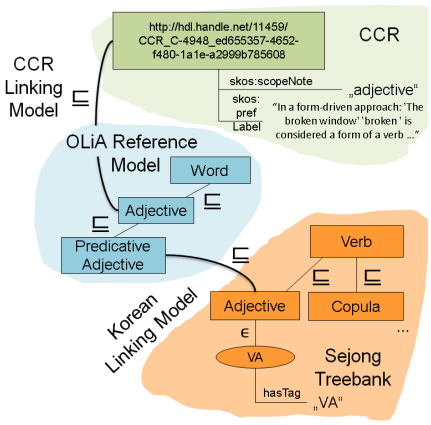

Figure 1: Modular OLiA ontologies

annotation model, a linking model defines rdfs:subClassOf-relationships between concepts in the annotation model and the reference model. Linking models are interpretations of annotation model concepts and properties in terms of the reference model. (4) Similarly, other community-maintained vocabularies are linked with OLiA, e. g., ISOCat and GOLD.

The OLiA ontologies cover different grammatical phenomena, including inflectional morphology, word classes, phrase and edge labels of different syntax annotations, as well as extensions for coreference, discourse relations, discourse structure and information structure (Chiarcos, 2014). Annotations for lexical semantics are only covered to the extent that they are found in syntactic and morphosyntactic annotation schemes.

3.2. Universal Part of Speech Tags (UP)

In order to facilitate annotation projection on the basis of parallel text, Petrov et al. (2012) introduced a highly reductionist, but 'universal' (i.e., cross-linguistically applicable, not universal in the sense of linguistic theory) representation of morphosyntactic annotation, based on a mapping of various annotation schemes to a minimal set of about a dozen part-of-speech tags. We integrate his information in OLiA by directly interpreting UP concepts as OLiA concepts and converting the mapping files from their original TSV format to OLiA annotation and linking models, using the following interpretation:

UP tag	UP definition	OLiA concept
VERB	verbs (all tenses and modes)	olia:Verb
NOUN	nouns (common and proper)	olia:Noun
PRON	pronouns	olia:Pronoun
ADJ	adjectives	olia:Adjective
ADV	adverbs	olia:Adverb
ADP	adpositions	olia:Adposition
CONJ	conjunctions	olia:Conjunction
DET	determiners	olia:Determiner
NUM	cardinal numbers	olia:CardinalNumber
PRT	particles or function words	olia:Particle
X	other	(unlinked)
.	punctuation	olia:Punctuation

The generated UP models are integrated into the stable release of OLiA in a separate directory `http://purl.org/olia/up`. On the one hand, this reflects their original source, but more importantly, the number of links that these linking models provide is much shallower than that of regular OLiA annotation models which aim to capture the full semantics of every tag. We thus distinguish regular linking models from UP linking models. Also, note that we did not include the UP linking models for annotation schemes previously covered by OLiA.

Note that UP tags are provided without definitions, there is thus no way to evaluate in a language-independent manner whether this mapping is correct. However, OLiA design encapsulates annotation models, so that OLiA users are free to replace an existing linking model with their own, e.g., by copying the linking model provided and adjusting it according to their specifications.

3.3. UD v.1

The Universal Dependencies comprise 'universal' part of speech annotations (originating in UP), morphosyntactic features (optional, originating from Zeman 2008, and, indirectly, from EAGLES) and syntactic dependencies (originating from the Stanford dependencies, De Marneffe and Manning 2008). UD data and documentation are available from `https://universaldependencies.org/` and have been subject to a long (and on-going) process of refinement that led to two iterations of the UD vocabulary (UD v.1 and UD v.2).

In relation to OLiA, which aims to cover all types of linguistic annotation, we see the UD community as working on a much more restricted problem, and on a limited (albeit impressive) scale of linguistic phenomena. We thus integrate UD specifications in the same way as an annotation model.

The UD ontology is, however, not compiled into a single file, but we provide three annotation models (one for parts of speech, morphosyntactic features, and dependency labels each), and their respective linking to OLiA. In these models, every 'universal' category (tag, feature value, dependency) is represented by a class, with its original documentation as an rdfs:comment, every language-specific category (tag, feature value, dependency) is represented as an instance of the corresponding class, again with its original documentation. Using an ontology browser such as Protégé[13], it is thus possible to quickly navigate from language-specific definitions to universal definitions, or to explore the range (and definitions) of language-specific instances of a universal category. All elements in the RDF graphs are identified by URIs that resolve to the respective page in the documentation, we thus complement the human-readable website with a conceptual layer that captures both its textual and structural content. Fig. 2 shows an example of the UD v.1 ontology for parts of speech.

Our UD ontologies are automatically derived from the documentation which resides in GitHub and is generated from Markdown using Jekyll.[14] We adopt the original URI

[13] `http://protege.stanford.edu`

[14] Our UD annotation models are automatically derived from the documentation by means of scripts, so that they revisions of the UD schema can be immediately reflected in the annotation

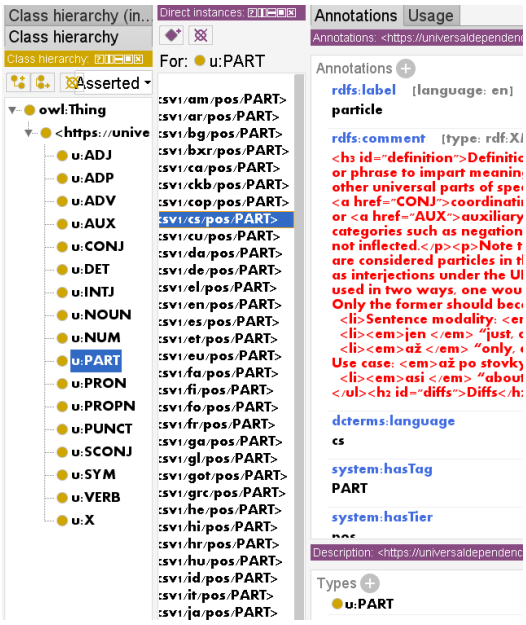

Figure 2: UD v.1 sub-ontology for parts of speech as visualized by Protégé

schema of the UD documentation, such that our URIs resolve to the original web page. For UD v.1, this was initially implemented using an experimental workflow that uses Jekyll templates to generate an RDF representation of the UD guidelines as part of the HTML rendering process:[15] Aside from mapping user-provided variables to specific HTML elements, we embedded RDF triples using the RDF in attributes specification (Herman et al., 2015, RDFa) to be able to read off an RDF representation directly from the website. As this RDF representation was generated from the original source, it was automatically synchronized with every modification of the website, and included all descriptive and all formal elements of the original website. As UD v.1 development is stalled, the UD v.1 specifications are now provided in a static fashion, we provide another RDF converter that now processes the generated HTML content, instead of the underlying Markdown.

For these ontologies, linking models have been created manually, where UD concepts were defined as subclasses of OLiA concepts. We only link universal categories for parts of speech, morphosyntactic features and dependency labels, language-specific extensions are not linked on an individual level, resp., only captured via their anchoring in universal categories.

The linking of parts of speech is largely equivalent with UP linking, the linking of dependency labels was based on the existing linking for Stanford dependencies. Neither of these required the introduction of novel OLiA concepts. As for morphosyntactic features, the situation is similar, as OLiA is partially based on the specifications of EAGLES which had a strong influence on the Interset inventory (Zeman, 2008) that represents the basis for UD feature annotations. The UD v.1 ontologies are included with OLiA[16] and, together with the build scripts, accessible from the OLiA GitHub repository.

3.4. UD v.2 and UniMorph

The revised UD v.2 vocabulary aims for establishing a greater level of coherence among the different layers of UD annotation and between different languages. It has been converted and linked analoguously to UD v.1 documentation, it should be noted, however, that the UD v.2 documentation is far less homogeneous than UD v.1 documentation, and that much language-specific information still refers to UD v.1. We provide the UD v.2 vocabularies and their linking with OLiA,[17] and the build scripts via the OLiA GitHub repository.

The Universal Morphology (UniMorph) project is a recent community effort aiming to complement the Universal Dependencies and their focus on syntax with coverage of inflectional morphology. UniMorph provides inflection tables for 110 languages using a TSV format, with lemma, form, and morphological features. Compatibility with UD is a requirement of the UniMorph community that has been partially achieved only (McCarthy et al., 2018), but by reference to a common reference vocabulary the relation between both vocabularies can be expressed easily. To this end, we created e a machine-readable representation of the UniMorph vocabulary in OWL and its linking with OLiA, closely following the approach taken for UD,[18] with build scripts included in the OLiA GitHub repository.

3.5. LexInfo

We focus on LexInfo v. 2.0, as this is closely coupled with the highly popular OntoLex-Lemon vocabulary and the reference vocabulary for lexical data categories in the Linguistic Linked Open Data cloud community.

Unlike UP, UD or UniMorph, LexInfo already comes as an ontology. We thus focus on its linking rather than its modelling choices. LexInfo is complementary to OLiA in the sense that OLiA takes a focus on annotations and the processing of natural language, whereas LexInfo provides an inventory of formal data categories for linguistic features of lexical entries and related information in dictionaries, wordnets and multilingual ontologies. So far, both ontologies have not been put into relation, although OLiA has been applied for encoding features of lexical resources, as well (Eckle-Kohler et al., 2015). Creating an interlinking between LexInfo and OLiA, and, via OLiA, with UniMorph and UD thus comes with the prospect of enormous synergies between lexical resources and natural language processing as well as for the enrichment of lexical resources and morphological resources, cf. Declerck and Racioppa (2019).

Similar to OLiA, LexInfo is partially based on ISOCat, but it differs from the OLiA Reference Model in

model. For a related piece of work, we would like to refer to Passos (2018) who provides a manually constructed and richer ontology for UD – which is, however, no longer synchronized with the ongoing improvement of UD documentation.

[15] This prototype is available via http://fginter.github.io/docs/.

[16] http://purl.org/olia/ud-v1

[17] purl.org/olia/ud-v2

[18] http://purl.org/olia/unimorph/

that it provides individuals rather than classes. Accordingly, it is not possible to establish formal equivalence relations (owl:equivalentClass), to assert identity (owl:sameAs) or to use the conventional OLiA linking properties (rdfs:subClassOf). Instead, LexInfo terms can only be defined as instances of OLiA concepts, so that LexInfo is linked with OLiA in the style of an OLiA annotation model,[19] i.e., a specialization for the domain of lexical information.

We provide a manual linking for LexInfo 2.0 with OLiA.[20] The linking is facilitated by the fact that LexInfo is based on ISOCat. No extensions of OLiA were necessary to represent LexInfo concepts. It is to be noted, however, that terminology relevant to the internal structure of lexical resources rather than their grammatical characteristics (another concern of the LMF model that which LexInfo developed from) have not been added to the OLiA Reference Model, but only information that could be potentially found in linguistic annotation.

3.6. CLARIN Concept Registry

Within the CLARIN infrastructure, the CLARIN Concept Registry (CCR) serves to provide semantically interoperable annotation terminology for various services and resources. It provides a collection of concepts, identifiable by persistent identifiers, that are relevant for the domain of language resources. At its core, the CLARIN CCR provides a revised subsection of ISOCat concept in SKOS, although under different URIs (handles).[21]

The actual RDF data is not publicly available, but can be recovered from the HTML rendering of the CCR browser. We provide a script that retrieves a partial RDF/Turtle representation for the morphosyntactic and syntactic facets of the CCR. In total, the morphosyntactic and syntactic facets of the CCR provide 484 terms, out of which 4 have been approved, 3 have been expired, and 477 (98.6%) remain at candidate status.

CCR terms are generally more (or, at least as) abstract than OLiA Reference Model terms. By analogy with the existing ISOCat linking (Chiarcos, 2010), we integrate CCR as an external reference model, i.e., to provide a linking that defines OLiA concepts as subclasses of (or equivalent classes with) CCR concepts. For technical reason, however, this is not possible with the native CCR data model: The RDF data drawn from the CCR portal natively represents terms as instances of skos:Concepts. In this form, CCR concepts are OWL instances (not classes) and can thus not serve as superclasses of OLiA concepts. Aside from parsing RDF triples out of the HTML, we thus perform OWL conversion as a post-processing step by replacing every skos:Concept with rdfs:Class, skos:broader with rdfs:subClassOf, etc., and add an OWL header. The original URIs are preserved and resolve to HTML pages in the CCR portal. After this conversion, OLiA concepts are manually be defined as subconcepts of the CCR ontology using rdfs:subClassOf in the linking model.

The original extracted CCR data, its ontological formalization and its linking is provided in the external branch of OLiA,[22] the crawler and conversion scripts are accessible via the OLiA GitHub repository.

4. Interoperability in Practice: Mapping Annotations

So far, we described the conversion and the linking of five post-ISOCat sources of linguistic reference terminology with OLiA, and thereby, with each other, with older resources such as ISOCat, GOLD, and the Typological Database System (TDS) ontology and a large number of annotation schemes. Each of these vocabularies is associated with a significant number of language resources that adhere to its specifications, and as a result, it is now possible to map annotations, resp., linguistic features of lexical resources from one vocabulary onto another.

It is to be noted, however, that such a mapping is not necessarily lossless: For a particular annotation scheme, linking with OLiA provides cross-linguistically applicable, intensionally defined concepts to formalize the meaning of tags. But OLiA does not necessarily capture resource-, tagset- or language-specific constraints that may apply to a particular tagset. This includes lexeme-specific tags, such as TO in the Penn Treebank tagset (for *to* in all its uses). In the linking, this is modelled as a subclass of olia:Preposition *or* (owl:unionOf) olia:Unique ('particle'), etc. While the possible functions of the elements tagged with TO are correctly captured in this way, we lose the information that the tag requires the presence of a particular word.

Another aspect where we encounter possible information loss between annotation models and OLiA are implicit constraints imposed by the underlying data structure of tagsets. Usually, this is a list or a tree, but in either case, the categories they posit are extensionally disjoint (in order to enable unambiguous tagging). In an ontology, they can overlap. If the underlying categories overlap, tagsets usually define what tag to be used. An attributive possessive pronoun like *her* in *her garden* is *both* a determiner (syntactically) and a pronoun (morphologically and semantically). If a tagset has a special tag for attributive possessive pronouns, it can be linked to the intersection of both classes (multiple inheritance), but if a tagset has only tags for pronoun and determiner (e.g., UD PRON and DET), a choice has to be made by the annotators or the tagset designers. In UD, for example, attributive possessive pronouns should be

[19] In this context, and given that other types of linguistic terminology resources have been linked in this way with the OLiA Reference Model (Dimitrova et al., 2016), it would be more adequate to use the term "domain model" rather than "annotation model": LexInfo is not concerned with annotations, but addresses domain-specific information for the domains of lexical and terminological resources (dictionaries, glossaries, word nets, terminologies, multilingual ontologies), where this information is actual content rather than an annotation attached to a content element. However, this comes with other connotations, so that we stay with the conventional term.

[20] `http://purl.org/olia/external/lexinfo`

[21] Via its public interface, the CLARIN CCR does not provide direct links to ISOCat concepts, but encodes them as comments in skos:changeNote literals. By means of regular expressions, the partial SKOS data, together with these ISOCat links have been retrieved.

[22] `http://purl.org/olia/external/ccr/`

tagged DET – even for languages without grammaticalized determiners. By linking DET with olia:Determiner, an attributive possessive pronoun in English would be correctly represented as determiner, but the annotation does not provide the information that it would be a pronoun, as well. When using OLiA to map from an annotation model that treats attributive possessive pronouns as pronouns to UD, we would not be able to predict the DET tag. When using OLiA for an annotation model that treats attributive possessive pronouns as a distinct class, OLiA would preserve the information that *her* is in both categories, but it would not be able to disambiguate the choice between DET *or* PRON in UD as a target tagset, because it is unaware of tagset-specific disambiguation rules.

As far as cross-linguistically applicable categories and features are concerned, we assume that an OLiA encoding for a particular tagset is lossless with respect to intensional semantics (if a particular category does not exist, it can be created), but that it can lose information about extensional restrictions. This is adequately expressed in the requirement to use of rdfs:subClassOf for the linking of annotation model concepts and the OLiA Reference Model. But this also means that the mapping from OLiA to an annotation model may be noisy, because extensional restrictions and disambiguation rules are not available.

The situation is different between OLiA and external reference models, e.g., concepts of CCR, GOLD, TDS or ISOCat, as we expect these to exist on the same or a higher level of abstraction as OLiA concepts. With the current linking, we do thus provide a lossless mapping from LexInfo, UP, UD, UniMorph and other annotation models via OLiA to CCR, GOLD and ISOCat. The mapping from CCR, GOLD, ISOCat or OLiA to UD and other annotation models, however, be lossy.

The overall relations of the vocabularies described here here are summarized in Fig. 3. For the reasons detailed above, all of them were integrated into the overall architecture of OLiA by means of rdfs:subClassOf (or, rdf:type, i.e., instance) relations. For UP, UD, and UniMorph, which are resource- or application-specific vocabularies, this is semantically adequate. LexInfo does provide a similar degree of abstraction as the OLiA Reference Model, but here, this modelling is required for technical reasons, i.e., that LexInfo individuals cannot be formalized as equivalent or superclasses of the OWL classes provided by the OLiA Reference Model. The situation is different for the CCR, which can be linked as an external reference model (that provides superclasses for OLiA concepts) after a conversion from SKOS to OWL. As for using rdfs:subClassOf relations to retrieve upper classes ('upward search'), our modelling assumes that this is lossless (if the linking is correct), as for using rdfs:subClassOf relations retrieve subclasses ('downward search'), our modelling assumes that this can be lossy (even if the linking is correct, we may miss extensional constraints).

With these vocabularies being linked, it is now possible to retrieve, say, the ISOCat concept for a given UniMorph annotation, say, the feature DAT (dative case) by upward search: In the UniMorph ontology, "DAT" is the value of the UniMorph label for unimorph:DAT, an instance of unimorph:DativalCase. Using SPARQL, we can retrieve the superclasses of this tag from the UniMorph graph:

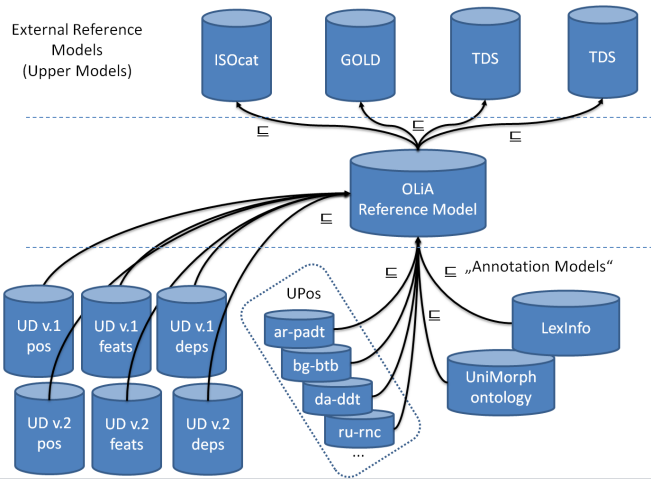

Figure 3: Linking of terminology repositories with each other, with ISOCat, GOLD and the TDS ontology via OLiA

```
SELECT ?uclass
WHERE {
 GRAPH <http://.../unimorph.owl> {
  ?uinst unimorph:label 'DAT'.
  ?uinst a/rdfs:subClassOf+ ?uclass. } }
```

The UniMorph linking can then be consulted to identify the corresponding OLiA concept, and from there, we can retrieve all ISOCat concepts.

```
SELECT ?uclass ?isoclass
WHERE {
 GRAPH <http://.../unimorph-link.rdf> {
  ?uclass rdfs:subClassOf ?oClass
 }
 GRAPH <http://.../olia.owl> {
  ?oClass rdfs:subClassOf* ?superClass
 } # *: direct or indirect superclasses
 GRAPH <http://.../dcr-link.rdf> {
  ?superClass rdfs:subClassOf* ?isoclass
 } }
```

From the OWL version of the ISOCat morphosyntactic profile provided by OLiA,[23] we can then retrieve the original datcat URIs for the corresponding classes:

```
SELECT ?isoclass ?uri
WHERE { ?isoclass dcr:datcat ?uri }
```

By navigating through the respective ontologies with SPARQL, we are thus able to map UniMorph "DAT" to `http://www.ISOCat.org/datcat/DC-3148`. This example shows an upward search, but analogously, search can be performed from ISOCat down to a specific annotation model, or from one annotation to another. At the moment, OLiA provides linkings with 94 annotation models for morphosyntax and/or syntax, applicable to more

[23] `http://purl.org/olia/external/dcr`

than 100 languages, the terminology repositories described here cover hundreds of language resources, as well, by linking them via the OLiA Reference Model, it is now possible to map each of these vocabularies to every other.

As the example shows, knowledge graphs, semantic technologies and ontologies are expressive and powerful devices, but they also come with considerable technological overhead for a relatively simple practical problem, the mapping of tags from one representation to another. More than the formal treatment of interoperability, this is of interest to the language resource community.

Using SPARQL SELECT, mapping tables can be easily generated with the queries above, as every variable binding of the SELECT statement will result in a row in a TSV, XML or JSON table, with one column per return variable. It is to be emphasized, however, that such mappings are not necessarily lossless or that they result in unambiguous results.

A suitable heuristic to eliminate redundant or too generic results is to restrict the results to elements on the shortest paths between two vocabulary elements in a SPARQL property path. In SPARQL, path length calculation can be implemented with aggregates:

```
SELECT ?x ?z (COUNT(?y) AS ?length)
WHERE {
 ?x rdfs:subClassOf* ?y.
 ?y rdfs:subClassOf+ ?z  }
```

Using a subsequent filtering step, mappings between two vocabularies can be restricted to pairs with minimal distance for every element of the source vocabulary.

5. Discussion and Outlook

In this paper, we describe the formalization of several linguistic annotation vocabularies by means of ontologies, and their respective linking with each other and existing terminology repositories by using the OLiA Reference Model as an intermediate representation. We provide the corresponding ontologies, resp. with their linking models, together with the Ontologies of Linguistic Annotations under a CC-BY license.

Grounding annotation schemes in formal ontologies, and interlinking them with each other establishes a high degree of interoperability at a comparably low cost. In particular, this approach differs from full-fledged standardization in that it does not require revisions of the actual annotation, but can be solely performed at the level of the vocabularies themselves.

From linked and formalized vocabularies, mapping tables that can be generated from the linking by means of SPARQL SELECT statements. As these involve queries over several RDF graphs, these are comparably complex. They do not, however, have to be developed by the end user. Instead, they can be statically compiled from a pair of vocabularies, and published along with software or data they are to be applied to. In general, the generated mapping rules will be approximative in the sense that they can be reductionistic (similar to the mappings provided by UP, the target annotations can be more coarse-grained) or underspecified (if the target vocabulary provides a higher degree of granularity than the source vocabulary, all alternative tags will be listed).

While it is not possible to guarantee 1:1 correspondences in mapping tables generated from such data (they might not even exist), this is nevertheless an efficient approach as it allows to retrieve approximative mapping rules for the transformation of annotations in accordance with dozens or even hundreds of language resources. With the linking of UP, UD, UniMorph, LexInfo, and the CCR, this functionality is now available for the most influential vocabularies for linguistic annotation terminology in the post-ISOCat era.

This paper is a resource description, in the sense that mappings between various vocabularies are being provided. At the same time, it employs linkings with the OLiA Reference Model as a basis for establishing a shared semantic space between them. While this seems to bring OLiA into a similar position as ISOCat and GOLD previously had (and for which they failed), we would like to emphasize that the point we are trying to make is something different: Using HTTP(S)-resolvable URIs for identifying concepts, semantically typed relations that hold between them (i.e., RDF properties), and the technical means to access remote data sets (i.e., RDF federation), it is possible to harmonize existing, distributed vocabularies. This is an insight that was already underling the development of the GOLD ontology. But in addition to that, we do not rely on centralized, monolithic repositories for data categories, application-specific interfaces and protocols, and formal means for concept registration and standardization by means of a designated commission or an editorial board as adopted by standardization-based approaches in the early 2000s. Instead, OLiA (like UD and, since January 2020, LexInfo) adopts a lean, software-inspired development workflow with a low entry barrier to its contributors: The vocabulary is maintained as an open source project and available via an open platform (GitHub). Privileged users (administrators) exist and they serve a similar role as the editorial board of ISOCat and GOLD, but *anyone* can fork a copy, modify it according to his needs – and request to merge his changes or additions back into the main branch.[24]

Our main contribution is thus to demonstrate the application of open source development principles and linked data technology to adress annotation interoperability challenges in a distributed setting: Concepts and definitions of different providers are defined in self-contained formal models (annotation models, terminology repositories) and can subsequently refer to vocabularies or reference concepts developed by a broader, and open community with a low entry barrier.

[24] It is very well conceivable that OLiA will be superseded by another terminology initiative at some point in time, but as long as it employs resolvable URIs for their concepts, these can be linked with OLiA and thus to all OLiA-linked vocabularies. But even if no explicit linking is provided: If this future vocabulary would be based on a fork of the OLiA Reference Model, and developed in a similar fashion, links with OLiA concepts will be recoverable from Diff/Merge scripts automatically created during version control and source code management.

Acknowledgements

The research described in this paper has been supported by BMBF Early Career Research Group 'Linked Open Dictionaries (LiODi)', the Horizon 2020 Research and Innovation Action 'Pret-a-LLOD', Grant Agreement number 825182, and the project 'Fachinformationsdienst Linguistik', funded by the German Research Foundation (DFG) in the infrastructure programme 'Wissenschaftliche Literaturversorgungs- und Informationssysteme'.

6. Bibliographical References

Brugman, H. and Lindeman, M. (2012). Publishing and exploiting vocabularies using the OpenSKOS. In *Proceedings of the Describing Language Resources with Metadata Workshop at LREC 2012*.

Chiarcos, C. and Sukhareva, M. (2015). OLiA - Ontologies of Linguistic Annotation. *Semantic Web Journal*, 518:379–386.

Chiarcos, C., Hellmann, S., and Nordhoff, S. (2012a). The Open Linguistics Working Group of the Open Knowledge Foundation. In *Linked Data in Linguistics*, pages 153–160. Springer, Heidelberg.

Chiarcos, C., Nordhoff, S., and Hellmann, S. (2012b). Interoperability of Corpora and Annotations. In C. Chiarcos, et al., editors, *Linked Data in Linguistics*, pages 161–179, Heidelberg, Germany. Springer.

Chiarcos, C. (2008). An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16.

Chiarcos, C. (2010). Grounding an ontology of linguistic annotations in the Data Category Registry. In *Proceedings of Language Resource and Language Technology Standards (LT&LTS) at LREC 2010*, pages 37–40, Valetta, Malta, May.

Chiarcos, C. (2014). Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 4569–4577.

Cimiano, P., Buitelaar, P., McCrae, J., and Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51.

Cimiano, P., McCrae, J., and Buitelaar, P. (2016). Lexicon Model for Ontologies. Technical report, W3C Community Report, 10 May 2016.

De Marneffe, M. and Manning, C. D. (2008). The Stanford typed dependencies representation. In *Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation at COLING 2008*, pages 1–8, Manchester, UK.

Declerck, T. and Racioppa, S. (2019). Porting multilingual morphological resources to OntoLex-Lemon. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP*.

Dimitriadis, A., Windhouwer, M., Saulwick, A., Goedemans, R., and Bíró, T. (2009). How to integrate databases without starting a typology war: The Typological Database System. In Martin Everaert, et al., editors, *The Use of Databases in Cross-Linguistic Studies*, Empirical Approaches to Language Typology [EALT] 41, page 155–208. Walter de Gruyter, Berlin.

Dimitrova, V., Fäth, C., Chiarcos, C., Renner-Westermann, H., and Abromeit, F. (2016). Building an ontological model of the BLL thesaurus. First steps towards an interface with the LLOD cloud. In *Proceedings of the Fifth Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources at LREC 2016*, page 50.

Eckle-Kohler, J., McCrae, J., and Chiarcos, C. (2015). *lemonUby* - A large, interlinked, syntactically-rich resource for ontologies. *Semantic Web Journal*, 6(4):371–378.

Farrar, S. and Langendoen, D. T. (2010). An OWL-DL implementation of GOLD: An ontology for the Semantic Web. In A. Witt et al., editors, *Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology*. Springer, Dordrecht, Netherlands.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. (2006). Lexical Markup Framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 233–236, Genoa, Italy.

Hahm, Y., Lim, K., Park, J., Yoon, Y., and Choi, K.-S. (2012). Korean NLP2RDF resources. In *Proceedings of the Tenth Workshop on Asian Language Resources (ALR 2012)*, pages 1–10, Mumbai, India.

Herman, I., Adida, B., Sporny, M., and Birbeck, M. (2015). RDFa 1.1 primer - third edition. W3C working group note, World Wide Web Consortium.

Ide, N. and Romary, L. (2004). A registry of standard data categories for linguistic annotation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 135–39, Lisboa, Portugal, May.

Kemps-Snijders, M. (2010). RELISH: Rendering endangered languages lexicons interoperable through standards harmonisation. In *7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages*, Valetta, Malta.

Leech, G. and Wilson, A. (1996). EAGLES recommendations for the morphosyntactic annotation of corpora. URL `http://www.ilc.cnr.it/EAGLES/annotate/annotate.html`. Version of March 1996.

McCarthy, A. D., Silfverberg, M., Cotterell, R., Hulden, M., and Yarowsky, D. (2018). Marrying universal dependencies and universal morphology. *arXiv preprint arXiv:1810.06743*.

McCrae, J. P., Labropoulou, P., Gracia, J., Villegas, M., Rodríguez-Doncel, V., and Cimiano, P. (2015). One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the web. In *European Semantic Web Conference*, pages 271–282. Springer.

Nivre, J., Agić, Ž., Ahrenberg, L., and et. al. (2016). Universal dependencies 1.4. http://hdl.handle.net/11234/1-1827.

Passos, G. P. (2018). *A formal specification for syntactic annotation and its usage in corpus development and maintenance: A case study in Uuniversal Ddependencies*. Ph.D. thesis, Universidade Federal do Rio de Janeiro.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2089–2096, Istanbul, Turkey.

Schmidt, T., Chiarcos, C., Lehmberg, T., Rehm, G., Witt, A., and Hinrichs, E. (2006). Avoiding data graveyards: From heterogeneous data collected in multiple research projects to sustainable linguistic resources. In *Proceedings of the E-MELD workshop on Digital Language Documentation*, East Lansing, Michigan, USA.

Schuurman, I. and Windhouwer, M. (2011). Explicit semantics for enriched documents. What do ISOcat, RELcat and SCHEMAcat have to offer? In *Proceedings of the 2nd Supporting Digital Humanities Conference (SDH 2011)*, Copenhagen, Denmark.

Schuurman, I., Windhouwer, M., Ohren, O., and Zeman, D. (2015). CLARIN Concept Registry: The new semantic registry. In *CLARIN 2015 Selected Papers*, pages 62–70.

Sylak-Glassman, J., Kirov, C., Yarowsky, D., and Que, R.). A Language-Independent Feature Schema for Inflectional Morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680. Association for Computational Linguistics.

Warburton, K. and Wright, S. (2020). A data category repository for language resources. In Antonio Pareja-Lora, et al., editors, *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*. MIT Press, Cambridge, Massachusetts.

Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.

Zinn, C., Hoppermann, C., and Trippel, T. (2012). The ISOcat registry reloaded. In *Proceedings of the Ninth Extended Semantic Web Conference (ESWC)*, pages 27–31, Heraklion, Greece.