Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 154–163
Marseille, 20-25 June 2022

Language Technologies for the Creation of Multilingual Terminologies. Lessons Learned from the SSHOC Project

Federica Gamba[1,2], **Francesca Frontini**[1,3], **Daan Broeder**[3], **Monica Monachini**[1]
[1]Istituto di Linguistica Computazionale "A. Zampolli" (ILC-CNR) Pisa, Italy,
[2]Charles University, Faculty of Mathematics and Physics, ÚFAL, Prague, [3]CLARIN ERIC
{federica.gamba, francesca.frontini, monica.monachini}@ilc.cnr.it, d.g.broeder@uu.nl

Abstract

This paper is framed in the context of the SSHOC project and aims at exploring how Language Technologies can help in promoting and facilitating multilingualism in the Social Sciences and Humanities (SSH). Although most SSH researchers produce culturally and societally relevant work in their local languages, metadata and vocabularies used in the SSH domain to describe and index research data are currently mostly in English. We thus investigate Natural Language Processing and Machine Translation approaches in view of providing resources and tools to foster multilingual access and discovery to SSH content across different languages. As case studies, we create and deliver as freely, openly available data a set of multilingual metadata concepts and an automatically extracted multilingual Data Stewardship terminology. The two case studies allow as well to evaluate performances of state-of-the-art tools and to derive a set of recommendations as to how best apply them. Although not adapted to the specific domain, the employed tools prove to be a valid asset to translation tasks. Nonetheless, validation of results by domain experts proficient in the language is an unavoidable phase of the whole workflow.

Keywords: Multilingual terminologies, data curation, language resource infrastructures

1. Introduction

The project Social Sciences and Humanities Open Cloud (SSHOC)[1] is one of the five disciplinary clusters funded under the INFRAEOSC 04-2018 call, whose mission is to connect the research infrastructures identified in the ESFRI Roadmap to the EOSC, the European Open Science Cloud[2].

In particular, SSHOC aims at realising the social sciences and humanities' part of EOSC. All SSH research infrastructures established as ESFRI Landmarks and Projects (CESSDA, CLARIN, DARIAH, ERIHS, ESS, SHARE), as well as relevant international SSH data infrastructures and the association of European research libraries (LIBER), participate in this project[3], and collaborate to integrate their data, tools, training materials, in an ecosystem that is in line with the FAIR principles (Ilijašić Veršić and Ausserhofer, 2019).

An important aspect for SSHOC is that of interoperable metadata and terminologies, which are crucial to ensure the discoverability of resources on the cloud. A common platform for publishing and sharing SSHOC vocabularies has been created to this purpose[4] and that should also serve the SSH community beyond the duration of the SSHOC project. An additional and important need within the SSH is the availability of high-quality multilingual vocabularies for discovery and other research tasks. In fact although English tends to be the dominant language of science, SSH researchers often produce culturally and socially relevant work in their local languages. While English can be used in metadata to classify such research outcomes, discovery could be greatly enhanced by the availability of descriptors in local languages.

CLARIN, the Common Language Resources and Technology Infrastructure[5], is a partner in SSHOC and is continuously exploring how Language Resources can contribute to create and sustain the SSH component of the EOSC (Broeder et al., 2020a) and how LT methods and practices can be adopted by already existing SSH research infrastructures to support their domain-specific work, as for instance in Broeder et al. (2020b), thus promoting and facilitating multilingualism in the SSH.

This paper relates to the outcomes of SSHOC on multilingual terminologies, led by ILC-CNR (which coordinates the Italian national node of CLARIN), in collaboration with CLARIN ERIC and CESSDA. The objective of the paper is many-fold: (i) investigating NLP and MT approaches in view of producing resources and tools to foster multilingual access to SSH content across different languages; (ii) assessing the performance of state-of-the-art technologies to this specific task; (iii) deriving indications for curators and infrastructure managers as to how best apply them, also taking into account the need for manual verification, as well as the best practices for publication. In particular, the paper will concentrate on two of the main case studies carried out in SSHOC: the evaluation of machine translation for metadata (described in 3) and the auto-

[1] https://www.sshopencloud.eu/.

[2] https://www.eosc.eu/.

[3] For more information about the SSHOC partners, cf. https://www.sshopencloud.eu/.

[4] https://vocabs.sshopencloud.eu/vocabularies/.

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

mated extraction of terms (described in 4). In SSHOC another case study was related to the creation of multilingual occupation ontologies, used in Social Science's surveys; in this paper, we will not report on the application of MT approaches to these specialised vocabularies and the reader is referred to Frontini et al. (2021) for the details.

2. State of the Art and Research Questions

Kulczycki et al. (2020) carried out a comprehensive study on multilingual publishing in the social sciences and humanities, analysing researchers in seven European countries. The results showed that, despite a great variability of practices and the general dominance of English, most SSH researchers produce culturally and societally relevant work in their local languages. For what concerns the creation of the SSH EOSC infrastructure part, this means that researchers in these domains may also need to be able search for research data and other resources by using non-English terms. Additionally, projects aiming at connecting publications and other research outcomes in the SSH will necessarily need a set of multilingual descriptors to ensure cross-lingual indexing and retrieval.

Nevertheless, metadata and vocabularies used in the SSH domain to describe and index research data are currently mostly in English. For instance the federated repositories of the CLARIN ERIC infrastructure mostly expose metadata in English, and the Virtual Language Observatory[6], CLARIN's meta catalogue, currently only allows for English searches. As to DARIAH ERIC, its vocabulary service[7] contains a number of terminologies, but only a limited amount of them are available in a language other than English. The domain of cultural heritage is probably the one where multilingual metadata and vocabularies is more widespread, and portals such as Ariadne Plus have put great effort into vocabulary mapping, so as to allow for multilingual search capabilities (Binding et al., 2019). In the social sciences the use of multilingual vocabularies is essential for comparative international studies where important social economic classifiers need to be used in surveys e.g. occupational titles, education levels etc. is a well-known important task and this requires high-level domain expert involvement.

When vocabularies in multiple languagages are not available or cannot be derived, however, the creation of multilingual metadata is indeed a time consuming effort, requiring experts with domain knowledge to translate not only the terms, but also the definitions to ensure a correct usage. With the progress of machine translation, it is thus important to assess the potential of current state of the art systems to assist in this process, and evaluate the amount of manual post correction necessary to obtain satisfactory results. Previous experiments in metadata translation showed promising results. For instance Song et al. (2020) assess the quality of metadata translation for *ukiyo-e* images from Japanese to English; however a limitation of that experiment is the use of the automatic evaluation metric BLEU, which requires existing manual translations for its calculation, and also will penalise potentially correct but divergent translations. Only manual assessment of translations by experts can thus provide a clear picture of the quality and usability of such systems, as will be shown in the next section.

At the same time, a number of areas and domains may not yet be sufficiently covered by pre-existing vocabularies. In such cases, the application of terminology extraction technologies can certainly help. Multililingual terminology extraction would be ideal, when applicable (Rigouts Terryn et al., 2020), but it requires the collection of comparable corpora covering all targeted languages, something which is not always possible. In the case study that was chosen for the SSHOC project, which focussed on the terminology around Data Stewardship, a sufficiently large corpus of relevant documents to extract candidate terms from could only be obtained for English. In such cases as these, monolingual terminology followed by machine translation can be considered as a viable alternative.

While the aforementioned language technologies are state of the art and are readily available, their application in the domain of data curation requires a thorough assessment of their costs and benefits, in order to ascertain how they can help maintain multilingual metadata and terminologies in a sustainable manner. In the following of sections (3 and 4) we shall describe the experiments carried out and the lessons learned, trying to answer respectively the following questions:

- Can Machine Translation (MT) tools offer an effective solution to translation tasks?
- How can Natural Language Processing (NLP) techniques and MT approaches help create new multilingual terminological resources?

Section 5 will present the produced datasets, which have been made publicly available via the CLARIN infrastructure and the SSHOC vocabularies platform.

3. Machine Translation for Metadata

As discussed in 2, metadata profiles are usually expressed in English, although the availability of multilingual metadata highly enhances the discoverability of datasets in the SSH. As a case study to investigate how MT can help address this need, we selected the metadata set of the CLARIN Concept Registry (CCR), which forms the basis of the semantic interoperability layer of CLARIN, especially as far as metadata are concerned[8]. From the CCR, the 232 metadata concepts

[6] https://vlo.clarin.eu.

[7] https://vocabs.dariah.eu/en/.

[8] To this end, it provides a collection of concepts that are each assigned a persistent identifier and a def-

classified as approved were selected; each metadata concept is always assigned a definition as well. In order to translate the approved set of metadata concepts and investigate the MT contribution, we employed various state-of-the-art tools: LINDAT Translation service[9] (Košarko et al., 2019), Deep-L[10], Google Translate[11] (Wu et al., 2016; Johnson et al., 2017), Reverso[12]. We selected these tools as they are easily available and exploitable as online services.

Our purpose was twofold: on the one hand, to obtain the translated set of 232 approved metadata concepts (Frontini et al., 2021b); on the other hand, to perform a qualitative assessment of state-of-the-art MT tools. By employing the four selected tools, we obtained an automatic translation of metadata concepts and their definitions into Dutch, French, Greek, and Italian, languages of the SSHOC WP3 partners and so allowing for result evaluation. Deep-L and Google Translate were employed for every language, whereas Reverso and LINDAT Translation service were exploited only in the case of covered languages (all except Greek in the case of Reverso; only French for LINDAT Translation service). Then, all the translations thus obtained underwent validation. Validators (one per language) were native/proficient speakers of the different languages, chosen based on their expertise on the topic. For each term, they assessed if its translation was correct (label 'yes'), partially correct ('maybe'), or incorrect ('no'). Similarly, for each definition they had to indicate whether the translation was substantially correct ('yes'), if it could get the general sense but some errors were present ('maybe'), or if it was substantially incorrect ('no'). The accuracy of automatic translations was later calculated by establishing the following criteria. If the translation of a term or a definition was validated as correct ('yes'), 1 point was assigned; if it was marked as partially correct ('maybe'), a score of 0.5 point was assigned, whereas in case of error ('no') the translation received 0 points. By adding up the scores thus obtained, a simple measure of the accuracy was returned.

The accuracy scores (see Tables 1, 2, 3, 4) highlight how Deep-L resulted to be the best MT tool among the tested ones, reaching the highest scores for each of the selected languages[13]. Google Translate returned good results and was always outperformed only by Deep-L. Therefore, Deep-L was employed as the preferred translation tool to define the translated metadata set, although some translations by Google Translate were retained if they validated better. Indeed, Deep-L not only obtains the best performances, but also has the maximum coverage as regards available languages. Moreover, Tables 1, 2, 3, 4 highlight a recurrent pattern in the performances of the tools: the obtained accuracy scores are always higher for definitions than for terms. This could be explained in two ways. On the one hand, the term is easier to translate if it is inserted in a wider context (i.e., the definition), since context contributes more elements and thus helps get the correct meaning, which is quite specific as technical concepts are concerned. This holds true in a few cases. However, most often the definition does not include the term: the better performances obtained with respect to definitions can therefore be explained by considering that the term itself has a very specific and technical meaning, whereas definitions mostly describe concepts by employing less-specific, thus easier to translate, words. Consider, for instance, the case of *planned*, understood as one of the possible values that can be used to describe an interview[14]: no French translation of the term (*prévu/prévue*) appropriately reflects its technical meaning (vs *planifié*), whereas its definition[15] is always translated correctly, regardless of the tool.

3.1. Lessons Learned

The selected case study highlights how the contribution of MT approaches to the creation of multilingual resources is of critical importance: the employed tools prove to be a valid asset to the translation task. It is important to underline that these tools were not adapted to the specific domains addressed by the chosen case study, and still they perform quite well. They clearly outperform traditional manual translation, as the decrease of translation quality is minimal compared to the gain in terms of time and effort needed. However, validation proves to be an unavoidable step when exploiting MT tools, which provide a solution to translation tasks but whose results need to be checked. Validation must be performed by domain experts, also having knowledge of the topic besides being proficient in the language.

4. Terminology Extraction of Data Stewardship Terminology

Together with multilingual metadata, multilingual terminologies fall within the scope of the creation of multilingual resources, aiming at facilitating knowledge discovery and classification and making content searchable across different languages. For the development of the European Open Science Cloud, terminolo-

inition. See https://www.clarin.eu/content/clarin-concept-registry.

[9] http://hdl.handle.net/11234/1-2922. Demo URL: https://lindat.mff.cuni.cz/services/translation.

[10] https://www.deepl.com/translator.

[11] https://translate.google.com.

[12] https://www.reverso.net.

[13] Note that Deep-L was also the best performing tool for japanese in Song et al. (2020).

[14] Cf. also *spontaneous*, *semi-spontaneous*, *elicited*.

[15] *The speaker prepares in detail the structure and content of his/her "performance" in advance* (en) > *L'orateur prépare en détail la structure et le contenu de sa "performance" à l'avance* (fr).

	Deep-L		Google Translate		Reverso	
	Term	Definition	Term	Definition	Term	Definition
Yes	175	223	147	209	144	214
Maybe	51	7	63	23	34	18
No	6	2	22	0	54	0
Total score	**200,5**	**226,5**	178,5	220,5	161	223
Score %	**86,42**	**97,63**	76,94	95,04	63,40	96,12

Table 1: Validation results of Dutch translations

	LINDAT		Deep-L		Google Translate		Reverso	
	Term	Definition	Term	Definition	Term	Definition	Term	Definition
Yes	184	204	197	217	195	212	189	208
Maybe	20	13	18	6	20	11	25	13
No	28	15	17	9	17	9	18	11
Total score	194	210,5	**206**	**220**	205	217,5	201,5	214,5
Score %	83,62	90,73	**88,79**	**94,83**	88,36	93,75	86,85	92,46

Table 2: Validation results of French translations

gies pertaining to data management are particularly important, as they can be used to enrich datasets descriptions but also other types of documentation.

We selected the topic of Data Curation and Stewardship as a case study to investigate how NLP techniques and MT approaches can help create new multilingual terminological resources. The intent was thus to use state-of-the-art language technologies to create a multilingual terminology specific to the domain of Data Stewardship. Such terminology (Frontini et al., 2021a), linked to other existing ones, will provide useful descriptors for datasets, but also, as stated by Molloy et al. (2021), could be used to create and assess Data Stewardship curricula, annotate FAIR-enabling training material, formalise job descriptions with competencies.

4.1. Corpus Creation

To begin with, we created a domain-specific corpus by collecting various documents pertaining to Data Stewardship and Curation. The corpus includes 70 open access documents among which standards and recommendations for Data Stewardship and Curation, deliverables, and other technical documents. All documents included are in English, and they amount to a total of 746,084 tokens. The sources of the documents are various: they were collected mostly from Research Data Alliance (RDA)[16] and through the OpenAIRE platform[17]. Since the chosen domain is recent and restricted, speaking of representativeness of the corpus is not accurate: although the process of finding the documents could not be exhaustive, all material that could be found was selected. However, as the domain of Data Stewardship is still expanding, in the future the corpus might need to be enlarged. A complete and detailed list of included documents is available in SSHOC D3.9 (Frontini et al., 2021).

4.2. Automatic Term Extraction

The second step consisted in automatic extraction of key terms. The intent was to obtain a preliminary list of terms relevant to the selected domain, to employ as a point of departure for the construction of the terminology, and to examine how Automatic Term Extraction (ATE) tools can contribute to the creation of similar resources. A study of the state-of-the-art in matter of terminology allowed us to select some tools suitable for corpus-based monolingual (English) extraction.

After a preliminary evaluation that took into consideration various tools, two of them were selected: TermoStat (Drouin, 2003) and TBXTools (Oliver and Vàzquez, 2015). The designed preliminary evaluation relied on an already annotated corpus, namely the ACTER dataset version 1.4 (Rigouts Terryn, 2020), and more specifically its English wind subcorpus, since no gold standard was available for our corpus. This preliminary evaluation did not mean to provide feedback on the state-of-the-art of tools for automatic term extraction, but intended to find a pragmatic solution to a precise task by meeting specific needs that strongly depend on the manual processing required to build the terminology. Therefore, we took into account not only raw scores of precision and recall, but also the amount of extracted terms. A brief description of tools and settings follows:

TermoStat (Drouin, 2003) performs ATE through the comparison of the focus corpus to a reference corpus, which in the case of English is a non-technical corpus encompassing articles from the Canadian daily newspaper The Gazette and excerpts from the

[16] https://www.rd-alliance.org/recommendations-and-outputs/catalogue.

[17] https://explore.openaire.eu.

	Deep-L		Google Translate	
	Term	Definition	Term	Definition
Yes	189	177	157	128
Maybe	14	38	17	65
No	29	17	58	39
Total score	**196**	**196**	165,5	160,5
Score %	**84,48**	**84,48**	71,34	69,18

Table 3: Validation results of Greek translations

	Deep-L		Google Translate		Reverso	
	Term	Definition	Term	Definition	Term	Definition
Yes	210	215	206	215	197	200
Maybe	12	12	11	11	21	23
No	10	5	15	6	14	9
Total score	**216**	**221**	211,5	220,5	207,5	211,5
Score %	**93,10**	**95,26**	91,16	95,04	89,44	91,16

Table 4: Validation results of Italian translations

British National Corpus (BNC). It extracts both simple terms and multi-word expressions. TermoStat first performs PoS tagging of the text thanks to the support of TreeTagger (Schmid, 1994). Thanks to a set of predefined syntactic matrices, term extraction is then performed. Every candidate receives a score based on the adopted statistical measure. After testing the four proposed measures (log-likelihood, log-odds ratio, specificity (Lafon, 1980), chi-square; we did not test raw frequency), we selected log likelihood, although it does not obtain substantially different results compared to the other statistical measures. No other parameters can be set (e.g., minimum frequency), so we manually excluded candidate terms occurring less than three times. The threshold of 3 is chosen as it represents an effective compromise between quantity and completeness, by allowing to sufficiently reduce the number of candidate terms, yet not excluding too many of them. TermoStat Web 3.0 was employed[18].

TBXTools (Oliver and Vàzquez, 2015) is a Python class which performs terminology extraction based on either a statistical or a linguistic approach.

- Statistical approach: we extracted up to trigrams and set the minimum frequency of candidates to be extracted to 3. We performed stop-word filtering, case normalisation, and nesting detection, which tries to spot shorter-term candidates that are not autonomous terms in and of themselves but are included in a longer term. A rejection-list of regular expressions allowed us to exclude listed patterns (mainly including non-word items). Candidates are then stored in a descending raw frequency order.

[18] `http://termostat.ling.umontreal.ca`.

- Linguistic approach: it requires a PoS-tagged corpus. TBXTools allows to lemmatise and PoS-tag a corpus by directly invoking the C++ library Freeling (Padró and Stanilovsky, 2012). Then, proper terminology extraction can be performed. A set of recurring PoS tags patterns allows to detect these same patterns in the corpus; the formalism for morpho-syntactic patterns allows as well to lemmatise the term candidates. We loaded a ready-made list of patterns for English, but patterns could have been automatically learnt with TBXTools from a tagged corpus and a set of known terms as well. As in the case of the statistical approach, we extracted unigrams, bigrams, and trigrams occurring at least 3 times. The extraction script already provided by TBXTools developers was slightly modified by adding case normalisation and nested normalisation, which proved useful during statistical extraction. Candidates are stored in a descending raw frequency order.

4.3. Validation

The extracted candidate terms were manually revised to remove undoubted errors and non-terms. All the terms selected as possibly correct were combined in a single list of 277 candidate terms, which underwent an external validation by domain experts, chosen from among the project partners and thus aware of the objectives of the task. Each of the two validators, separately to avoid any potential reciprocal conditioning, was asked the question "Is the term, as used in the example, a specific term of the domain of Data Stewardship?".
We compared validation results by defining the following criteria. In case of agreement, no issues arose: if validators agreed in considering a term valid (answer 'yes'), the term was kept, whereas in case of an agreed 'no' the term was discarded. When a validator an-

swered 'yes' to the question and the other answered 'maybe', 'yes' prevailed; conversely, in case of 'no' and 'maybe', 'no' prevailed. In case of 'maybe' agreement, as well as when a validator did not consider the term as valid while the other did, disagreement resolution was necessary. This was resolved by evaluating if the term was already included in other terminologies, as well as by looking for definitions in other resources or in the corpus. For instance, Validator 1 validated the term *big data*, but this was not validated by Validator 2: since a definition could be found in the corpus, it was kept as a valid term. Conversely, in the case of *data author* Validator 1 answered 'no', whereas Validator 2 selected 'yes': the term was eventually discarded, as it did not occur in any other terminology and a valid definition could not be found as well. The final list of validated terms encompasses 260 entries.

We selected linearly weighted Cohen's k as a measure of Inter-Annotator Agreement (IAA), and obtained a 0.08 Cohen's k value, corresponding to a slight agreement according to the classification in Artstein and Poesio (2008). The low result obtained is indicative of a still low standardisation of terminology pertaining to the selected domain of Data Stewardship. For this reason, the resulting terminology was obtained after analysing divergent cases with the two validators. However it may need to undergo further discussion, as Data Stewardship terminology is still evolving and needs to be stabilised.

The final list of validated terms also constituted the gold standard to employ to evaluate the accuracy of the tools. Obviously, such a gold standard cannot be considered exhaustive, as it does not include all terms occurring in the corpus, but could still serve as a reference to evaluate tools. Precision, recall and F-score were calculated for each tool (see Table 5).

	TBXTools statistical	TBXTools linguistic	TermoStat
Terms	6582	3742	3789
Precision	4.53%	5.29%	8.50%
Recall	89.39%	73.48%	83.71%
F-score	8.66%	9.87%	15.43%

Table 5: Accuracy scores for TermoStat and TBXTools

Overall, TermoStat shows the best balance between the number of extracted terms and the extraction accuracy, as proven by the F-score. The assessment of precision and recall followed two slightly different criteria. As far as precision is concerned, all variants of the gold standard terms were considered correct: for instance, if a system extracted *data center*, *data centers*, *data centre*, *data centres*, all the four expressions were counted as correctly extracted terms, in order not to penalise the tools that did not perform lemmatisation. As for recall, of course all variants of a same term could not be counted as true positives, otherwise the number of correctly extracted terms (true positives) would have exceeded the total number of terms in the gold standard (true positives + false negatives). For this reason, in the above-mentioned example of *data center* all the four possible forms of the term (*data centre*, *data centers*, *data centres*) were counted as one entry while calculating recall.

4.4. Definitions, Linking and Translation

After validation some terms, which represented different labels referring to a same concept, were merged into one single entry (concept), yet referred to by multiple labels. For instance, *data citation* and *citation of data* were considered as pointing at the same concept, to which the verbal equivalent *cite data* was assigned as well. As a result, the Multilingual Data Stewardship Terminology consists of 211 distinct concepts. Concepts[19] were then provided with definitions derived from different sources: other terminologies, if the term was there found and defined; the corpus itself; papers or Web articles. When no definition for a term could be found in any of these sources, a new definition was written[20]. Our approach was mainly intensional. We created definitions consistent with those we derived from other sources; in particular, we tried to be as consistent as possible with Loterre's approach. However, in the case of the creation of a fully-fledged terminological resource, it would be necessary to align all definitions with respect to a same approach (intensional/extensional). The ease with which definitions for terms were found correlates with the degree of standardisation of the term: for some terms the definition was easier to find, and such terms turned out to be more standardised within the domain of interest. For instance, for a common and standardised term like *interoperability*, multiple definitions were found. Moreover, some terms are borrowed from the Information and Communication Technology (ICT) domain, thus holding an already high degree of standardisation.

Besides assigning a definition, for each term it was also verified if it occurs in other existing terminologies: Loterre Open Science Thesaurus [21] (47 matching terms), Linked Open Vocabularies (LOV) platform[22] (42 matching terms), terms4FAIRskills[23] (65 matching terms). ISO Online Browsing Platform (OBP)[24] allows for the querying of terms defined in ISO standardisation documents and was consulted as well, although

[19] Specificities could exist in the different languages, thus requiring a definition at term level, instead of concept level. However, for our purposes this was not taken into account.

[20] E.g. *discovery metadata*: *metadata that are used for the discovery of data, often in the context of data archives.*

[21] Developed at Inist-CNRS; see `https://www.loterre.fr/skosmos/TSO`.

[22] `https://lov.linkeddata.es/dataset/lov`.

[23] `https://github.com/terms4fairskills/FAIRterminology/tree/master/initial_prototyping`.

[24] `https://www.iso.org/obp/ui/#search`.

not systematically; if a corresponding entry was found, it was linked with the term at hand (89 times).
Each pair of terms and definitions was then translated into multiple languages. We decided to automatically translate the collected terms and definitions with DeepL, since it resulted as the best performing MT tool among the ones tested with respect to Multilingual Metadata (see Section 3). Selected languages, that are the languages of the WP3 partners, are: Dutch, French, German, Greek, Italian, Slovenian. Translations underwent an external validation by native speakers, in order to correct inaccurate translations.

4.5. Lessons Learned

The workflow described so far with respect to the creation of the Multilingual Data Stewardship Terminology represents a valid methodology that can be adopted every time a new (multilingual) terminology is to be created. Indeed, validation of translations required about one week. The limited amount of employed time is promising in terms of sustainability and scalability: if necessary, the Data Stewardship Terminology can easily and rapidly grow and include more languages, as long as a native speaker with some domain expertise is available for validating automatic translations. The same holds true for potentially any other terminology, whether already multilingual but prone to include more languages, or monolingual and intending to evolve into a multilingual one. As observed in section 3.1, validation is unavoidable when exploiting NLP strategies (in this case, MT and ATE), which however prove to be valid assets to similar tasks.
Overall, the intention was not to create a fully-fledged terminology, with a solid hierarchical structure, given that current initiatives in this sense already exist: what was done can be seen as a contribution, which added relevant terms to existing resources while showing how automatic translation tools can help in similar tasks.

5. Publishing the Results

The created terminologies were then converted and made available in SKOS[25], as it is the recommended format discussed in D3.1 (Broeder et al., 2019) and the underlying model of the SKOSMOS Vocabulary publication platform[26] (Monachini et al., 2021).
The SKOSification, i.e. transforming the flat tabular vocabularies that were used as a work-format, was carried out with the support of a conversion tool developed at ISTI-CNR[27]. The Data Stewardship Terminology and Multilingual Metadata are ingested in the mapper as spreadsheets; the mapper parses the spreadsheets and transforms the content in SKOS data by applying a set of mapping rules. The result of the mapping is an RDF Graph, which is formatted according to the Terse RDF Triple Language (Turtle) data format and finally stored in two separate files.

With respect to the Multilingual Data Stewardship Terminology, every concept is assigned a unique subject identifier, a `prefLabel` for each language (English, Dutch, French, German, Greek, Italian, Slovenian). If present, alternative forms are expressed through `skos:altLabel` property and are tagged based on the language in which they are formulated. The `altLabel` property allows not only to encode synonyms (e.g., *data representation - representation of data*) and acronyms (e.g., *Digital Object Identifier - DO*), but it also provides a solution for handling alternative spelling variants (e.g., *anonymisation - anonymization*), often due to differences between UK and US English. The representation of spelling variants represents one of the challenges that are related to multilinguality. Among these, were cases where distinct `prefLabel` and `altLabel` in English (e.g., *data cleaning - data cleansing*) had an identical translation in another language (e.g., in Italian both terms are translated as *pulizia dei dati*). Similar cases were handled by conflating the identical translations into one unique translation, considered as a `prefLabel`. Each concept is also assigned a definition, whose source is reported as well. Linking to other existing terminologies when appropriate was performed through the `skos:exactMatch` property. This was possible in case of linking to Loterre or resources from the Linked Open Vocabularies (LOV). However, in case of ISO norms and terms in terms4FAIRskills, a linking through `skos:exactMatch` was not possible since terms within ISO norms are not identified through a URI, and neither terms in terms4FAIRskills are assigned a proper one, as the resource is still under development. Therefore, when linking with one of these resources was possible, the `skos:note` property was used, whose object is a literal and does not require a valid URI. No internal hierarchy was defined, but a shallow one was provided by linking the concepts extracted to broader terms in other terminologies, if possible. The solution is not ideal: in the future a better integration of such interlinked resources should be achieved. This work was just intended as a case study to test tools and a methodology, but in order to develop a fully-fledged terminology hierarchy should be addressed deeply.

As regards Multilingual Metadata, a similar approach was adopted. For each entry a `prefLabel`, a definition, and a source of the definition are specified. All `prefLabel`s and definitions are available in multiple languages (English, Dutch, French, Greek, Italian). Each metadata term is then linked to the corresponding persistent identifier in the CCR through the `skos:exactMatch` property.

[25] The use of SKOS and the publication on the SSHOC terminological platform were requirements of the project, but other models of term representation are possible.

[26] `http://www.skosmos.org`.

[27] The mapper, implemented in a Python Notebook, is now available at `http://hdl.handle.net/20.500.11752/ILC-566` and will be published in the SSH Open MarketPlace.

Resource	Question	Results	Recommendations
Multilingual Metadata	Can MT tools offer an effective solution to translation tasks?	MT tools perform well, although their results need to undergo validation.	Promote community collaboration to encourage vocabulary reuse, avoid duplication of efforts and further test NLP techniques and MT approaches.
Multilingual Data Stewardship Terminology	How can NLP techniques and MT approaches help create new multilingual terminological resources?	ATE and MT make a significant contribution to the creation of multilingual resources. Yet, results need to be checked: domain experts, also having knowledge of the topic, are necessary for validation.	

Table 6: Results and remarks

SKOS proved to have a sufficient degree of expressivity for what concerns the Multilingual Metadata concepts and the Multilingual Data Stewardship Terminology. More complex vocabularies are not needed to encode similar structures.
Both Multilingual Metadata[28] and the Multilingual Data Stewardship Terminology[29] are freely and openly available through CLARIN and SSHOC infrastructures (VLO and SSHOC vocabularies platform, respectively) under the CC BY 4.0 license.

6. Recap and Recommendations

The presented case studies allow to derive a first set of recommendations (see Table 6) which can be addressed to the SSH community at large, but most and most specifically, to the research infrastructures that are part of SSHOC and that will maintain the planned continued SSH colaborations after the lifetime of the project. A few considerations are in order. Concerning hierarchisation, Multilingual Metadata concepts have been provided in SKOS format, in the form of a flat list. They will have to be integrated with metadata schema and the associated vocabularies, such as for instance the CCR. In such contexts, the associated vocabularies usually already provide concept hierarchies and should be respected. Therefore, the Multilingual Data Stewardship Terminology is provided without a full-fledged hierarchy allowing more easy integration; a partial hierarchy is obtained through linking it to other existing terminologies such as Loterre Open Science Thesaurus and terms4FAIRskill. However, these last resources as well are not yet finalised and not stable enough, making it premature to simply link the terms extracted to them and consider the task to be done. In the future, community collaboration at SSHOC and EOSC level should be promoted, to encourage vocabulary reuse, avoid duplication of efforts and further test the extraction and translation approaches adopted. Indeed, the contribution of NLP approaches and MT tools to the creation of multilingual resources is crucial, representing valuable resources for translation tasks. Nonetheless, validation of results is an unavoidable phase of the whole workflow, and should be carried out by domain experts who are also proficient in the language.

7. Conclusion and Future Work

In this paper we outlined the results of activities in the SSHOC project, aimed at assessing the usability of Language Technologies for data curation in SSH, with two case studies on metadata translation and terminology creation. The tools that have been employed proved to be a valid asset to translation tasks. It is important to underline that these tools are not adapted to the specific domains addressed by the chosen case studies, and still they perform quite well. These promising results lead us to believe that Language Technologies can become a very useful tool for research infrastructures, especially in the SSH, to support multilingual description of data thus improving their findability.
At the same time, the current technologies are not without limitations and cannot completely replace manual curation. For this reason any duplication of efforts should be avoided at all costs. A first important step in this direction was the creation of a common terminological platform, which will facilitate finding, translating and reusing existing vocabularies. Other important steps will involve the collaboration with initiatives such as the EOSC Task Forces[30] will in particular for what concerns the terminology around Data Stewardship and Data Curation, which will allow the correct description of Open Data practices throughout Europe, thus facilitating the implementation of a common Open Science agenda. The active participation of CLARIN members in EOSC related activities is a step in this direction.

8. Acknowledgements

The work reported here has received funding from the EU H2020 research and innovation programme (g.a.

[28] The dataset is available at `http://hdl.handle.net/20.500.11752/ILC-568` and at `https://vocabs.sshopencloud.eu/vocabularies/sshocmm`.

[29] The dataset is available at `http://hdl.handle.net/20.500.11752/ILC-567` and at `https://vocabs.sshopencloud.eu/vocabularies/sshocterm`.

[30] `https://www.eosc.eu/advisory-groups`.

823782) for project SSHOC and has been supported by CLARIN-ERIC and CLARIN-IT. We gratefully acknowledge support from Charles University, grant No. SVV 260 575. We thank LINDAT/CLARIAH-CZ for their support.

9. Bibliographical References

Artstein, R. and Poesio, M. (2008). Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Binding, C., Tudhope, D., and Vlachidis, A. (2019). A study of semantic integration across archaeological data and reports in different languages. *Journal of Information Science*, 45(3):364–386, June. Publisher: SAGE Publications Ltd.

Broeder, D., Trippel, T., Degl'Innocenti, E., Giacomi, R., Sanesi, M., Kleemola, M., Moilanen, K., Ala-Lahti, H., Jordan, C., Alfredsson, I., L'Hours, H., and Ďurčo, M. (2019). SSHOC D3.1 Report on SSHOC (meta)data interoperability problems, December. Final version approved by the European Commission on 18 November 2019.

Broeder, D., Eskevich, M., and Monachini, M. (2020a). LR4SSHOC: The Future of Language Resources in the Context of the Social Sciences and Humanities Open Cloud. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, pages 33–36, Marseille, France, May. European Language Resources Association.

Broeder, D., Eskevich, M., and Monachini, M. (2020b). *Proceedings of the Workshop about Language Resources for the SSH Cloud*. Marseille, France, May. European Language Resources Association.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Frontini, F., Gamba, F., Monachini, M., Broeder, D., Tijdens, K., and Vipavc Brvar, I. (2021). SSHOC D3.9 Report on Ontology and Vocabulary Collection and Publication.

Ilijašić Veršić, I. and Ausserhofer, J. (2019). Die sozial- und geisteswissenschaften und ihre interoperabilität mit der european open science cloud: Was ist sshoc? *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, 72(2):383–391, Dez.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Košarko, O., Variš, D., and Popel, M. (2019). LINDAT translation service. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kulczycki, E., Guns, R., Pölönen, J., Engels, T. C. E., Rozkosz, E. A., Zuccala, A. A., Bruun, K., Eskola, O., Starčič, A. I., Petr, M., and Sivertsen, G. (2020). Multilingual publishing in the social sciences and humanities: A seven-country European study. *Journal of the Association for Information Science and Technology*, 71(11).

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1(1):127–165.

Molloy, L., McQuilton, P., and Le Franc, Y. (2021). EOSC Co-creation funded project 074: Delivery of a proof of concept for terms4FAIRskills: Technical report, March.

Monachini, M., Jääskeläinen, T., Uytvanck, D. V., der Lek, I. V., Broeder, D., and Moranville, Y. (2021). MS8 Choice of Vocabulary Publication platform for SSHOC, August.

Oliver, A. and Vàzquez, M. (2015). Tbxtools: A free, fast and flexible tool for automatic terminology extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 473–479.

Padró, L. and Stanilovsky, E. (2012). FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2473–2479, Istanbul, Turkey, May.

Rigouts Terryn, A., Hoste, V., and Lefever, E. (2020). In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation*, 54(2):385–418, June.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Song, Y., Batjargal, B., and Maeda, A. (2020). A preliminary attempt to evaluate machine translations of ukiyo-e metadata records. In Emi Ishita, et al., editors, *Digital Libraries at Times of Massive Societal Transition*, pages 262–268, Cham. Springer International Publishing.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

10. Language Resource References

Frontini, F., and Gamba, F., and Monachini, M., and Broeder, D. (2021a). *SSHOC Multilingual Data Stewardship Terminology*.

Frontini, F., and Gamba, F., and Monachini, M., and Broeder, D. (2021b). *SSHOC Multilingual Metadata*.

Rigouts Terryn, A. (2020). *ACTER (Annotated Corpora for Term Extraction Research) v1.4*.