Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782 Start Date of Project: 01/01/2019 Duration: 40 months

Deliverable 3.9 Report on Ontology and Vocabulary Collection and Publication

Dissemination Level	PU
Due Date of Deliverable	31/12/2021 (M36)
Actual Submission Date	20/12/2021
Work Package	WP3 - Lifting Technologies and Services into the SSH Cloud
Task	Task 3.1 Multilingual Terminology
Type	Report
Approval Status	Waiting EC approval
Version	V1.0
Number of Pages	p. 1 – p. 45

Abstract:

The deliverable presents three detailed case studies for each of the main topical areas of SSHOC Task 3.1 “Multilingual Terminologies'' aiming to investigate NLP and MT approaches in view of providing resources and tools to foster multilingual access to SSH content across different languages and improve discovery by non-native speakers. A set of multilingual metadata concepts, multilingual vocabularies and automatically extracted multilingual terminologies has been delivered as freely, openly available data, fully corresponding to the FAIR principles promoted within the EOSC, findable through the VLO and other CLARIN and SSHOC services.

SSHOC, “Social Sciences and Humanities Open Cloud”, has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020); H2020-INFRAEOSC-04-2018, under the agreement No. 823782

History

Version	Date	Reason	Revised by
0.0	22/12/2021	Discussion among co-authors	All authors
0.1	07/12/2021	Task leader send to peer reviewers	Mari Kleemola (CESSDA/FSD) Nicolas Larrousse, Edward Gray (HUMA-NUM)
0.2	10/12/2021	Address internal review comments	CNR-ILC and CLARIN
1.0	17/12/2021	Final version for submission	

Author List

Organisation	Name	Contact Information
CNR-ILC	Francesca Frontini, Federica Gamba, Monica Monachini	francesca.frontini@ilc.cnr.it federica.gamba@ilc.cnr.it monica.monachini@ilc.cnr.it
CLARIN/ERIC	Daan Broeder	d.g.broeder@uu.nl
CLARIN/WWI	Kea Tijdens	K.G.Tijdens@uva.nl
CESSDA/UL-ADP	Irena Vipavc Brvar	irena.vipavc@fdv.uni-lj.si

Contributor List

Organisation	Name	Contact Information
CESSDA/UL-ADP	Luka Opršnik	luka.opresnik@fdv.uni-lj.si
FSD Finnish Social Science Data Archive	Taina Jääskeläinen	taina.jaaskelainen@tuni.fi
CNR-ISTI	Cesare Concordia Luca Trupiano	cesare.concordia@isti.cnr.it, luca.trupiano@isti.cnr.it
CLARIN/Athena	Maria Gavrilidou	maria@ilsp.athena-innovation.gr
UL/FDV	Janez Štebe	Janez.Stebe@fdv.uni-lj.si
FORS	Christina Bornatici	christina.bornatici@fors.unil.ch
CLARIN/CUNI	Dušan Variš	varis@ufal.mff.cuni.cz

Table of Contents

Executive Summary

This deliverable pertains to SSHOC Task 3.1 which was responsible for investigating and providing resources and tools to support the multilingual aspects of the future pan-EU SSH infrastructure.

Making data and services accessible and usable in SSH is very much also a matter of providing relevant translations, translation of metadata concepts, multilingual vocabularies, terminology extraction across languages, multilingual databases.

The deliverable offers a detailed report on the gathering and translation of relevant SSH metadata, ontologies and vocabularies for the use-cases indicated in the task's topics: multilingual metadata concepts and vocabularies, the multilingual occupation ontology, with cross-country female occupational titles.

In accordance with SSHOC and the EOSC FAIR recommendations and requirements, the metadata vocabularies and ontologies have been published via several different formats and facilities.

Section 1. The introduction sets the landscape and describes the need of multilingual vocabularies both for classification and discovery in the context of a cloud-based infrastructure that will offer access to research data and related services adapted to the needs of the SSH community.

Section 2. "Multilingual metadata" investigates the possibility to use and test Natural Language Processing (NLP) approaches and Machine Translation (MT) to make the metadata more accessible using national languages other than English. A selected case study was the recommended metadata set of the CLARIN Concept Registry (CCR): the whole set of metadata and definitions were translated into French, Greek, and Italian. The section describes the machine-translation and evaluation process, also comparing different technologies.

Section 3. "Multilingual vocabularies and ontologies" introduces two other typical case-studies. The first one addresses one of the pressing needs in social sciences research. Many surveys, indeed, ask respondents to specify their occupation and the occupational ontology is used for the survey questions. For many languages the occupational titles for males and females are not identical. In section 3.1 the enrichment of the occupational ontology with lists for male and female titles, is described for many languages, namely for Dutch, German, Slovenian and French.

The second case study focuses on the automatic extraction of terminology from texts: a list of domain-specific terms was automatically extracted from a corpus of Data Curation and Stewardship, validated by domain experts, automatically translated into multiple languages (Dutch, French, German, Greek, Italian, Slovenian) and linked to other existing terminologies.

Section 4. describes the SKOS-ification and publication process of the results, together with the challenges posed by multilinguality.

Section 5. offers an overview of the exploitation and sustainability of the results and how these are made available to the community.

Finally the Conclusions provide some reflections on Machine Translation approaches adopted for translating the vocabularies into multiple languages, the advantages in terms of time saving and some first recommendations to the community.

Abbreviations and Acronyms

ACTER	Annotated Corpora for Term Extraction Research
ADP	Arhiv Družboslovnih Podatkov - Slovenian Social Science Data Archives
API	Application Programming Interface
BNC	British National Corpus
CCR	CLARIN Concept Registry
CESSDA	Consortium of European Social Science Data Archives
CLARIN	Common Language Resources and Technology Infrastructure
CMDI	Component MetaData Infrastructure
CNR	Consiglio Nazionale delle Ricerche - Italian National Research Council
CNRS	Centre National de la Recherche Scientifique
CUNI	Charles University – Prague
DLC	Data Life Cycle
EOSC	European Open Science Cloud
ERIC	European Research Infrastructure Consortium
ESCO	European Skills/Competences, qualifications and Occupations
FAIR	Findability Accessibility Interoperability Reusability
FDV	Fakulteta za družbene vede -- Faculty of Social Sciences
FORS	Swiss Centre of Expertise in the Social Sciences
FSD	Finnish Social Science Data Archive
IAA	Inter-Annotator Agreement
ICT	Information and Communication Technology
ICTeSSH	International Conference on ICT enhanced Social Sciences and Humanities
IETF	Internet Engineering Task Force
ILC	Istituto di Linguistica Computazionale - Institute for Computational Linguistics

INIST	Institut de l'Information Scientifique et Technique
ISCO	International Standard Classification of Occupations
ISO OBP	International Standard Organization Online Browsing Platform
ISTI	Istituto di Scienza e Tecnologie dell'Informazione - Institute of Information Science and Technologies
LOST	Loterre Open Science Thesaurus
LOV	Linked Open Vocabularies
LSTM	Long Short-Term Memory
MP	Market Place
MT	Machine Translation
MWE	Multi-Word Expression
NLP	Natural Language Processing
OeAW	Österreichische Akademie der Wissenschaften
POS	Part-of-Speech
RDA	Research Data Alliance
RDF	Resource Description Framework
SKOS	Simple Knowledge Organization System
SKP	Slovenska klasifikacija poklicev – Slovene classification of occupation
SSH	Social Science and Humanities
SSHOC	Social Science and Humanities Open Cloud
UFAL	Institute of Formal and Applied Linguistics
UI	User Interface
URI	Uniform Resource Identifier
VLO	Virtual Language Observatory
WWI	WageIndicator
XKOS	Extended Knowledge Organization System

1. Introduction

Resource discovery and classification are an important part of the Data Life Cycle (DLC) and using the appropriate vocabularies can greatly improve both discovery and classification. Consequently, for SSHOC it is important to address this issue with respect to the SSH domain.

An additional need within the SSH is that of high-quality multilingual vocabularies. In fact, as shown by Kulczycki et al (2020), 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.

Language technologies can diminish the cost of creating and maintaining multilingual vocabularies. SSHOC Task 3.1 "Multilingual Terminology", led by CNR-ILC with contribution by the partners CLARIN/ERIC, CLARIN/WWI, CLARIN/Athena, CLARIN/CUNI, CESSDA/FSD, CESSDA/ADP, is investigating and providing tools and resources to support the multilingual aspects of the pan-European SSH infrastructure.

Making data and services accessible and usable in SSH is also very much a matter of providing relevant translations, translation of metadata concepts, multilingual vocabularies, terminology extraction across languages, and multilingual databases. In line with these main objectives, in this task, hence, several vocabularies have been created that help facilitate resource discovery and classification. Part of the creation process was done using and testing NLP approaches and machine translation, to make the vocabularies more useful in environments with localised user-interfaces (UI) and (strong) requirements for using national languages other than English.

- Multilingual metadata concepts (lead CNR-ILC and partners CLARIN/ERIC, CLARIN/Athena, CLARIN/CUNI, CESSDA/FSD): the 232 approved metadata concepts from the CLARIN Concept Registry (CCR), together with their definition, were collected and translated into multiple languages (Dutch, French, Greek, Italian) by exploiting different MT tools (cf. Section 2. below)
- Multilingual occupation ontology (lead CLARIN/WWI, CESSDA): The multilingual occupation ontology, developed in T3.2 "Selected SSH Ontologies and Vocabularies", consists of translated occupational titles from the list compiled by the WageIndicator Foundation[1]; for these titles male and female forms in different languages were provided and translations were checked by labour market experts.
- Data Stewardship Multilingual Terminology (lead CNR-ILC with partner CLARIN/ERIC, CLARIN/CUNI, CESSDA/ADP, CESSDA/FDV, CESSDA/FORS): a list of 210 domain-specific concepts was automatically extracted from a corpus of Data Curation and Stewardship, validated by

[1] WageIndicator Foundation Occupational titles list: https://wageindicator.org/Wageindicatorfoundation; [14 December 2021]. Surveycodings: https://www.surveycodings.org/articles/home; [14 December 2021].

domain experts, automatically translated into multiple languages (Dutch, French, German, Greek, Italian, Slovenian) and linked to other existing terminologies.

For the topical use-cases existing vocabularies, or implicitly existing vocabularies -- i.e. extracted from existing corpora -- were selected for testing, rather than new vocabularies that are under construction, since the foreseen procedures require existing language models and evidence for translation and evaluation.

Due to the strong focus on vocabularies, this Task launched and contributed to the "Vocabulary Initiative", an initiative that has emerged from the need to align the vocabulary activities across the SSHOC work packages and to optimise the sharing of research data across various practices and domains. A common vocabulary approach was also the topic of the SSHOC ICTeSSH workshop jointly organised by this Task and other project's Tasks[2] (see Monachini, 2020).

The main outcomes of the virtual information and discussion sessions organised by the Vocabulary Initiative confirmed that: in general vocabularies are essential for SSH resource classification and discovery and in SSHOC context are essential for the SSH Open Marketplace (MP) to describe the entries, improve search and retrieval, and foster discoverability; vocabularies should be published as Linked Open Data based on Simple Knowledge Organization System (SKOS)[3] data model and provide comprehensive coverage of the domain through concept definitions; users should be able to reuse existing vocabularies or link them to other artefacts, thus ensuring semantic interoperability. Finally, yet importantly, since vocabularies are changing once they are integrated into a platform, updates and maintenance should be ensured and done systematically. This will ensure quality.

In accordance with SSHOC and the EOSC FAIR recommendations and requirements, the results have been or will be published shortly via a number of different formats and facilities:

- in the ILC4CLARIN centre repository
    - SSHOC Multilingual Metadata http://hdl.handle.net/20.500.11752/ILC-568
    - SSHOC Multilingual Data-Stewardship Terminology http://hdl.handle.net/20.500.11752/ILC-567
- as part of the SSHOC results vocabularies on the SSH Vocabulary Commons platform at OEAW http://vocabs.sshopencloud.eu/vocabularies [from start 2022]
- findable via the SSH Open Marketplace

Task 3.1 collaborates in SSHOC WP3 with task 3.2 "Selected SSH Ontologies and Vocabularies" in view of the common topic of multilingual vocabularies and with all SSHOC activities in general that have an interest in (multilingual) vocabularies, through its SSH Vocabulary Initiative activities.

[2] SSHOC workshop at ICTeSSH conference proceedings: https://ictessh.pubpub.org/pub/hskk7vmx/release/2; [14 December 2021].
[3] SKOS reference guide: https://www.w3.org/TR/skos-reference/; [14 December 2021].

2. Multilingual Metadata

As argued in the Introduction, the availability of multilingual metadata highly enhances the discoverability of datasets in the SSH. However, their creation implies translations from English (the language normally used in metadata profiles) into various languages. This is a cumbersome task, and metadata modellers and curators may not have the linguistic skills to produce good translations in some languages. Machine translation may be a solution, but a thorough evaluation of the quality of the results provided by the various state of the art systems is necessary.

Within the context of the topical area "creating Multilingual metadata and taxonomies for discovery", the selected case study was the metadata set of the CLARIN Concept Registry (CCR)[4]. The CCR forms the basis of the semantic interoperability layer of CLARIN, especially as far as metadata are concerned (cf. Component MetaData Infrastructure - CMDI). To this end, it provides a collection of concepts that are each assigned a persistent identifier. From the CCR, the 232 metadata concepts which are classified as approved (*status: approved*) were selected[5]. In the CCR, each metadata concept is also assigned a definition.

As the goal was to obtain a multilingual set of metadata concepts, the next step consisted in translating the approved set of metadata concepts. Therefore, in order to translate the metadata concepts and their definitions, the LINDAT Translation service (Košarko et al., 2019)[6], i.e. the Machine Translation tool developed at UFAL, Charles University in Prague, was employed. Indeed, the tool was made available to the partners of the SSHOC project and was employed as well in SSHOC Task 4.2 "Preparing tools for the use of Computer Assisted Translation" so as to translate into different languages the answers of surveys. The LINDAT Translation service, a neural networks-based translation service, provides a simple UI and API that allows the use of pre-trained Transformer models served by TensorFlow Serving[7].

However, the provided tool covers only seven languages, namely Czech, English, French, German, Hindi, Polish, and Russian. Thus, in order to translate metadata concepts into more and different languages, it was decided to employ other state-of-the-art tools as well: Deep-L[8], Google Translate[9], and Reverso[10]. Deep-L working principles have not been disclosed; it is only stated that it exploits artificial neural networks, as in the case of Reverso. Google Translate is a Neural Machine Translation system consisting

[4] CLARIN Concept Registry: https://www.clarin.eu/content/clarin-concept-registry; [14 December 2021]. Accessible by read-only facetted browser at https://concepts.clarin.eu/ccr/browser/; [14 December 2021].
[5] CLARIN uses two levels of CCR concept evaluation and curation, and the approved concepts have been discussed and agreed upon by the CLARIN CCR board, this reflects the CCR status Nov. 2021.
[6] LINDAT Translation Service: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2922; [14 December 2021]. Demo URL: https://lindat.mff.cuni.cz/services/translation; [14 December 2021].
[7] TensorFlow Serving: https://www.tensorflow.org/tfx/guide/serving; [14 December 2021].
[8] Deep-L: https://www.deepl.com/translator; [14 December 2021].
[9] Google Translate: https://translate.google.com/; [14 December 2021].
[10] Reverso: https://www.reverso.net/; [14 December 2021].

of a deep LSTM network with 8 encoder and 8 decoder layers using residual connections as well as attention connections from the decoder network to the encoder (see Wu et al., 2016; Johnson et al., 2017). Thanks to the support of the four selected tools, an automatic translation of the whole set of metadata and definitions was obtained. Metadata concepts and definitions were translated into Dutch, French, Greek, and Italian (that are the languages of the SSHOC WP partners). Deep-L and Google Translate were employed for every language, whereas Reverso and the LINDAT Translation service were exploited only in the case of covered languages (all but Greek in the case of Reverso; only French as for the LINDAT Translation service).

After obtaining the automatic translations as described, it was necessary for all translations to be validated by native or proficient speakers of the different languages. Validators were chosen based on their expertise on the topic.- First, validators performed an evaluation of the automatic translations (see Figure 1). For each term, they assessed if its translation was correct (label '*yes*'), partially correct (label '*maybe*') or incorrect (label '*no*'). Similarly, for each definition they had to indicate whether the translation was substantially correct (label '*yes*'), if it could get the general sense but some errors were present (label '*maybe*') or if it was substantially incorrect (label '*no*'). Optionally, validators also had the chance to rank the systems based on their accuracy, from the best performing to the worst performing one.

		UFAL			DEEP-L (https://www.deepl.com/en/tra			Google translate (https://translate.goo			REVERSO			List better system	
term	definition	sou	term transl	te definition translation (UFAL)	de	definition val	ter definition tra	de	term translation	ter	definition t	de	term translatio	terr definition t	definitio	term (U,D,G, R)	definition (U,D,G, R)
wizard-of-oz	A research experiment i	(soi	magicien de	N Expérience de recherche dans laquel	Y	magicien d'oz	N Une expérienc	Y	magicien d'oz	N	Une expérie	Y	magicien de l'oz	N Une expérier	Y	//	D,G,R
writing systems	The visual representatic	(soi	systèmes d'é	Y La représentation visuelle de la langu	Y	systèmes d'écri	Y La représentat	Y	systèmes d'écriture	Y	La représent	Y	systèmes d'écritu	Y La représent	Y	D,G,R	D,G,R
persistent identifi	Specification of a persis	(soi	identifiant pe	M Spécification d'un identificateur pers	M	identifiant pers	M Spécification c	Y	identifiant persistant	M	Spécification	M	identifiant persis	M Spécification	Y	//	D,R
public	The access to the comm	(soi	public	Y L'accès à l'événement de communica	Y	public	Y L'accès à l'évé	Y	publique	Y	L'accès à l'év	Y	public	Y L'accès à l'év	Y	D,G,R	D,G,R
address	The address of an organ	(soi	adresse	Y L'adresse d'une organisation qui a pa	Y	adresse	Y L'adresse d'un	Y	adresse	Y	L'adresse d'u	Y	adresse	Y Adresse d'une	Y	D,G,R	D,G,R
age	The number of years thi	(soi	âge	Y Le nombre d'années de vie de quelqu	Y	âge	Y Le nombre d'a	Y	âge	Y	Le nombre d	Y	âge	Y Le nombre d	Y	D,G,R	D,G,R
elicited	Investigator asks speake	(soi	élimé	N L'enquêteur demande au ou aux locu	Y	a suscité	M L'enquêteur di	Y	suscité	M	L'enquêteur	Y	obtenue	M L'enquêteur	Y	//	D,G,R
email	The email address of a p	(soi	courriel	M L'adresse électronique d'une personn	Y	email	Y L'adresse élec	Y	email	Y	L'adresse e-i	Y	courriel	M L'adresse co	Y	D,G	D,G,R
event structure	Indicates the structure (	(soi	structure évé	Y Indique la structure de l'événement (	Y	structure des é	M Indique la stru	Y	structure de l'événer	M	Indique la st	Y	structure d'évène	M Indique la sti	Y	//	D,G,R
execution locatio	Identification of the loc	(soi	lieu d'exécut	Y Identification de l'endroit où l'outil o	Y	lieu d'exécutio	Y Identification (	Y	lieu d'exécution	Y	Identificatio	Y	emplacement d'e	M Identificatior	Y	D,G	D,G,R
experimental sett	A transmission of the cc	(soi	cadre expéri	Y Une transmission du contenu se déro	Y	milieu expérim	M Une transmiss	Y	cadre expérimental	Y	Une transmi	Y	cadre expérimen	Y Transmissior	Y	G,R	D,G,R
face to face	The transmission of the	(soi	face à face	Y La transmission du message assure u	Y	face à face	Y La transmissio	Y	face à face	Y	La transmiss	Y	face à face	Y La transmiss	Y	D,G,R	D,G,R
false	Contrary to what is true	(soi	faux	Y Contrairement à ce qui est vrai, erroi	Y	faux	Y Contraire à ce	Y	faux	Y	Contraireme	Y	faux	Y Contraireme	Y	D,G,R	D,G,R
family	The access to the comm	(soi	famille	Y L'accès à l'événement de communica	Y	famille	Y L'accès à l'évé	Y	famille	Y	L'accès à l'év	Y	famille	Y L'accès à l'év	Y	D,G,R	D,G,R
fax number	The Fax number of a pe	(soi	numéro de t	M Le numéro de télécopieur d'une pers	M	numéro de fax	Y Le numéro de	Y	numéro de fax	Y	Le numéro c	Y	numéro de téléc	M Numéro de t	M	D,G	D,G
monologue	Communication event w	(soi	monologue	Y Événement de communication avec (	Y	monologue	Y Événement de	Y	monologue	Y	Événement (	Y	monologue	Y Événement c	Y	D,G,R	D,G,R
morphology	The study of the structu	(soi	morphologie	Y L'étude de la structure et de l'élector	N	morphologie	Y L'étude de la s	Y	morphologie	Y	L'étude de la	Y	morphologie	Y L'étude de la	Y	D,G,R	D,G,R
planned	The speaker prepares in	(soi	prévu	N L'intervenant prépare en détail la str	Y	prévu	N L'orateur prép	Y	prévue	N	L'orateur pr	Y	prévue	N L'orateur pré	Y	//	D,G,R
planning type	Indicates in how far the	(soi	type de plan	Y Indique dans quelle mesure le consu	Y	type de planific	Y Indique dans c	Y	type de planification	Y	Indique dan	Y	type de planificat	Y Indique dans	Y	D,G,R	D,G,R

Figure 1. The case of French as an example of the validation spreadsheet, which contains the original terms and definitions, their translations, and columns devoted to validation and ranking of the systems.

The accuracy of automatic translations has been 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 (label '*no*') the translation received 0 points. By adding up the scores thus obtained (see 'Total score' in the tables below), a simple measure of the accuracy was returned. Tables 1, 2, 3 and 4 report the results and highlight in bold the best performances.

	Deep-L		Google Translate		Reverso	
	Terms	Definitions	Terms	Definitions	Terms	Definitions
N of yes	175	223	147	209	144	214
N of maybe	51	7	63	23	34	18
N of no	6	2	22	0	54	0
Total = max score	232	232	232	232	232	232
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 Translation		Deep-L		Google Translate		Reverso	
	Terms	Definitions	Terms	Definitions	Terms	Definitions	Terms	Definitions
N of yes	184	204	197	217	195	212	189	208
N of maybe	20	13	18	6	20	11	25	13
N of no	28	15	17	9	17	9	18	11
Total = max score	232	232	232	232	232	232	232	232
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.

	Deep-L		Google Translate	
	Terms	Definitions	Terms	Definitions
N of yes	189	177	157	128
N of maybe	14	38	17	65
N of no	29	17	58	39
Total = max score	232	232	232	232
Total score	**196**	**195**	165,5	160,5
Score %	**84,48%**	**84,05%**	71,34%	69,18%

Table 3. Validation results of Greek translations.

	Deep-L		Google Translate		Reverso	
	Terms	Definitions	Terms	Definitions	Terms	Definitions
N of yes	210	215	206	215	197	200
N of maybe	12	12	11	11	21	23
N of no	10	5	15	6	14	9
Total = max score	232	232	232	232	232	232
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.

In light of the accuracy scores, it can be observed how Deep-L resulted to be the best MT tool among the tested ones, reaching the highest scores for each of the selected languages (Dutch, French, Greek, Italian). Google Translate returned good results and was always outperformed only by Deep-L. So Deep-L was employed as the preferred translation tool, although some translations by Google Translate were retained if they validated better. Therefore, Deep-L was selected as the most efficient tool and thus employed in Topic 3.1.2 as well (see Section 3.2), since not only it obtains the best performances, but it also has the maximum coverage as regards available languages.

Moreover, the tables above 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.

The dataset is available at: SSHOC Multilingual Metadata - http://hdl.handle.net/20.500.11752/ILC-568.

3. Multilingual Vocabularies and Ontologies

3.1 Multilingual occupation ontology

3.1.1 Description of the ontology

A first topic that Task 3.1 focussed on is the multilingual occupation ontology. Occupation is a key variable in socio-economic research because occupations are equally important for an individual's identity as for their working life, earnings capacity, social life, friendships, and social status.

In many social sciences surveys, respondents are asked 'What is your occupation?'. For decades, the answer to this question was registered in a text box, followed by data entry of the answers, and a subsequent coding process. This procedure is expensive and time consuming. A survey of 1,000 respondents could easily comprise 500 different occupational titles.

Statistical offices and survey agencies developed their own national coding procedures, according to their own, national occupational classification. It was only in the 2000s that survey holders aimed to facilitate cross-country comparisons by coding the occupational titles into the international ISCO-08 classification (ILO 2012). Although this International Standard Classification of Occupations (ISCO) is widely used and it is the standard in the European Union, quite a few countries stick to their national classifications and at best apply mapping tables to the ISCO-08 classification.

With an increasing internationalisation of the economies of many countries and an increasing cross-national mobility of individuals, a cross-country comparison of occupational titles has gained importance. This has challenged the cross-national reliability of coding, which is important when drafting conclusions about occupational careers, occupational earnings, occupational entry levels,

occupational certification, occupational boundaries, and other labour market features in a European or wider context (e.g., Meng et al, 2020).

Two issues needed to be solved regarding the measurement of occupations in labour market studies: the costs of the coding process had to be reduced and the international comparability of occupational classification had to be improved. In the next paragraphs it will be explained how these two problems were solved by developing a multilingual, multinational occupational ontology, including ISCO-08 codes for all titles in the master list of the ontology.

With the increasing use of web-based surveys the open text box for the 'What is your occupation' question could be replaced by a list of occupational titles, facilitating respondents to self-select the occupational title most applicable to their job title. Since the early 2000s WageIndicator Foundation (WWI) developed such a list for its continuous web-survey on work and wages in the Netherlands. All titles in the list were coded according to the ISCO classification. From 2004 on, WageIndicator could expand its web survey to another seven European Union countries, and in the following years to countries outside Europe. Again, the choice was in favour of a predefined list and not in favour of post-coding text strings, due to budgetary considerations. This predefined list was derived from the titles listed in the ISCO-08 manual (ILO, 2012) and expanded with titles proposed by WageIndicator partners. All titles in this source list were coded according to the ISCO-08 classification. To populate the database of occupations for new countries and new languages, project partners were asked to translate the source list and add new titles if they thought they were missing. So, the number of titles in the source list grew, and so did the number of translations (Tijdens, 2015). Currently, the database consists of more than 4,200 titles (Tijdens, 2019a). The latest ontology is posted on the website[11].

The ontology consists of translated occupational titles of the master list. Given that occupational titles are an entity of national labour markets, the titles can be different in countries with the same language, though they still refer to the same tasks in the occupation. For example the *Domestic help* is called *Huishoudelijke hulp particulieren* or *Werkster* in the Netherlands and *Poetsvrouw* in Flemish for Belgium. For this reason, the translated titles are country specific, and identifiable by their locales, thus by their language and country, according to the IETF BCP 47 language tag followed by the ISO 3166-1 alpha-2 country code, in this example nl_NL and nl_BE. Given the problem of national occupational titles, a literal translation of occupational titles often does not provide a correct occupational title for the country at stake. Thus, translations should rather be provided by national labour market experts than by translators. The main criterion to judge whether an occupational title in one language has the same meaning as an occupational title in another language is whether the job holders at stake perform the same tasks. Though job descriptions and task lists are available at the 4-digit level of the ISCO classification, the occupational titles in the ontology are disaggregated to a 5-digit level to ease survey respondents' self-identification. Unfortunately, detailed job descriptions and task lists are often

[11] Surveycodings: https://www.surveycodings.org/articles/home; [14 December 2021].

missing, implying that the role of the labour market expert in the translation process is even more important.

The national characteristics of occupational titles may explain why machine translation performs poorly. Using Google Translate the translations of the English source list were compared to the translations available in the ontology. In the four countries compared, none of the percentages of correct translations was above 20%, as Table 5 shows. For this reason the use of MT for translating occupational titles was not further explored in this use-case. The four countries were selected because the occupational titles in the database were recently reviewed by national experts, as part of the SSHOC activities.

	Finnish (fi_FI)	Italian (it_IT)	Dutch (nl_NL)	Slovenian (sl_SI)
corresponding percentage	17%	20%	14%	16%
N of tested titles	551	498	1009	57

Table 5. Validation results of translations for Finland, Italy, Netherlands and Slovenia, showing the percentage of corresponding translations from Google Translate and the ontology.

As part of SSHOC task 3.1, SSHOC partners from Finland, Switzerland and Slovenia checked the translations in the ontology and proposed improvements. In the next section, the work of the Slovenian partner ADP is presented below as a use case, because this partner not only contributed by improving the database, but also prepared a document explaining what hurdles they came across.

3.1.2 Use case: Slovenia

SSHOC partner ADP was tasked to review and upgrade the current list of occupations and broader terms for the Slovenian language as available through the SurveyCodings tool in light of matching items with official occupation titles used in Slovenia. From the primary foreseen quick review and creation of female form of the code, this ended up as being an extremely complex task, mainly because some of the occupational titles translated in Slovenian did not reflect the occupational title commonly used in this country.

Detailed review showed that many of the terms currently listed are not used as such in day to day work of Slovene companies. After consultation with the Slovenian Statistical Office, it was decided to use ISCO08 standard, the SKP-08 standard[12], and the ESCO Occupations database[13] as a base. There was

[12] SKP – Slovenska klasifikacija poklicev – Slovene classification of occupation: https://www.stat.si/skp/; [14 December 2021].
[13] ESCO – European Skills/Competences, qualifications and Occupations: https://ec.europa.eu/esco/portal/occupation; [14 December 2021].

consultation of Termania[14] (online dictionary of technical terms), various government websites, Official gazettes of RS, terminological database of public relations[15], Slovene Literary Language Dictionary[16] and experts in the fields of public relation and communication, international relations and defence.

At this point, 248 items were either successfully confirmed or a new phrase was proposed. Of these, there are 57 occupations, to which a female version of the occupation was added. The rest are categories. Additional explanations are provided for items of more complex nature when consultation with different terminology or occupation databases took place. This additional information is provideto Kea Tijdens, the SSHOC partner responsible for the database.

Main issues that occur while cleaning existing database are listed:

- When the occupation listed does not exist in Slovenia and there is no single person who performs this function, but there does exist an official Slovenian translation. Such cases were "governor" and "traditional chief or head of village". Since there is no sense in having such occupations in the Slovenian database, they are marked with a slash «/», to indicate an empty field.
- Where the occupation translated in Slovenian does not exist in the same form. In such cases, the field was filled with the occupation most adjacent in meaning. Such a case was "fire commissioner".
- Some cases were translated into the same word in Slovenian since there is no difference between the occupations in Slovenian formal organisation. Such were "embassy representative", "ambassador", "diplomatic representative", and "diplomat", which was resolved by filling in translation only for one element.

In the process of confirming translations, other issues with the database were noticed. Categories listed under codes ranging from -20500 to -20509 were not listed in the live database, which made them only harder to translate since no context existed. Additionally, a bug with visual presentations of the results in the live database was identified. When opening submenus in which the list of occupations is really long (example being “Menedžement, upravljanje>Najvišja stopnja menedžmenta v organizaciji z manj kot 50 zaposlenim") only part of the list is visible on the page. Dynamic sizing of the page could be implemented.

ADP plans to review a minimum 200 additional elements by the end of January 2022, for which additional public databases or acts will be used. Main two being the online Catalogue of functions, jobs,

[14] Termania: https://www.termania.net/; [14 December 2021].
[15] Terminological database of public relations: https://www.termania.net/slovarji/termis-terminoloska-podatkovna-zbirka-odnosov-z-javnostmi/7967666/publiciteta?ld=111&query=Publiciteta&SearchIn=Linked; [14 December 2021].
[16] Slovene Literary Language Dictionary: https://fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika; [14 December 2021].

and titles in public sector[17], which combines decrees from several public administration sectors, and Public Sector Salary System Act[7][18].

Matching and translation has proved to be a challenging task, which requires serious consideration, planning and budget to complete. A future project involving terminological experts would be necessary to complete the task. Scope notes or descriptions of an occupation, as well availability of alternative titles would be extremely beneficial in the time of translation. Examples of this approach can be seen in the ESCO database.

3.1.3 Gendered occupational titles

In some countries, male and female occupational titles are different, whereas in other countries only gender-neutral or male occupational titles are in use. For example, for Germany, the English word DTP operator is translated to DTP Operator for men and DTP Operatorin for women. Most German occupational titles have distinct words for males and females, challenging the use of the word DTP Operator/in in the ontology. However, this hampers respondent's easy readability and thus understanding when selecting the proper occupational title in the results of the text string matching. In web surveys it is easy to include a so-called gender filter so that male and female respondents can search in the list for their gender only. If an occupation has a female title different from its male title, the female title will be shown, otherwise the male or neutral titles will be shown.

For four languages the percentage of gender-distinct occupational titles was explored, namely for Finland, the Netherlands, Germany, and Slovenia. Finland reported that none of the occupational titles for men and women are different between the genders since Finnish has no gender markers in its language. In the Netherlands at maximum 31% of the occupational titles could be phrased differently for the two genders. However, the use of female titles for women is not yet settled, because in day-to-day communication increasingly the male version of the occupation is used for women. For example, the word *medewerker* refers to men and *medewerkster* refers to women, though increasingly the word *medewerker* is used for both genders. The Delpher archive of Dutch newspapers provides 1.098.392 hits for *medewerker* and only 241.421 for *medewerkster* for all nation-wide newspapers in the 21st century (18%). In Germany, in contrast, the majority of occupational titles are phrased differently for male and female jobholders and sensitivity to gendered phrasing seems to be widely present in Germany. In the ontology, slightly over 75% of the titles are phrased differently. Using the female titles from the Slovenian partner, a similar percentage is noticed, namely 72%.

[17] A Catalogue of functions, jobs and titles in the public sector: http://www.pportal.gov.si/FDMN/index.html; [14 December 2021].
[18] Public Sector Salary System Act: http://www.pisrs.si/Pis.web/pregledPredpisa?id=ZAKO3328; [14 December 2021].

3.2 Multilingual terminology: case study on Data Stewardship

Another topical area in Task 3.1 focuses on extraction of terminology from technical documentation about standards and interoperability. 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, terminologies pertaining to data management are particularly important, as they can be used to enrich datasets descriptions but also other types of documentation. As proof of this, in February 2021 the EOSC Co-creation funded project delivered a proof of concept terminology[19] capturing the skills and competencies necessary to make and keep data FAIR, so as to enable the cross domain and cross-repository searching for training materials by the skills and competencies they require and confer (see the Report on terms4FAIRskills by Molloy et al. 2021). Later in the same year, the EOSC Task Force in Data Stewardship Curricula and Career Paths[20] was launched; among its core activities and goals is the clarification of terminology around Data Stewards/Data Stewardship (Task 1).

For these reasons, the topic of Data Curation and Stewardship, which is of the utmost importance to all research infrastructures operating within the framework of the EOSC, was selected. 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, linked to other existing ones (cf. below 3.2.4) will provide useful descriptors for datasets, but also, as indicated by the Report on terms4FAIRskills, could be used to create and assess data stewardship curricula, annotate FAIR-enabling training material, formalise job descriptions with competencies.

To pursue this goal, a terminology extraction from technical documentation about standards and interoperability was performed, in order to integrate the extracted, validated and translated terms in existing lexical-semantic/terminological/ontological resources, with the goal of providing background resources to be used for the basic access functionalities. Moreover, this approach allows investigating how and to what extent language technologies (in this specific case, tools for automatic term extraction and machine translation) can assist in the creation of domain-specific terminologies.
The workflow to create the Data Stewardship Multilingual Terminology is shown by the diagram below (Figure 2) and hereafter illustrated in detail.

[19] terms4FAIRSkills: https://github.com/terms4fairskills/FAIRterminology/tree/master/initial_prototyping; [14 December 2021].
[20] EOSCTask Force in Data Stewardship Curricula and Career Paths: https://www.eosc.eu/sites/default/files/tfcharters/eosca_tfdatastewardshipcurriculaandcareerpaths_draftcharter_20210614.pdf; [14 December 2021].

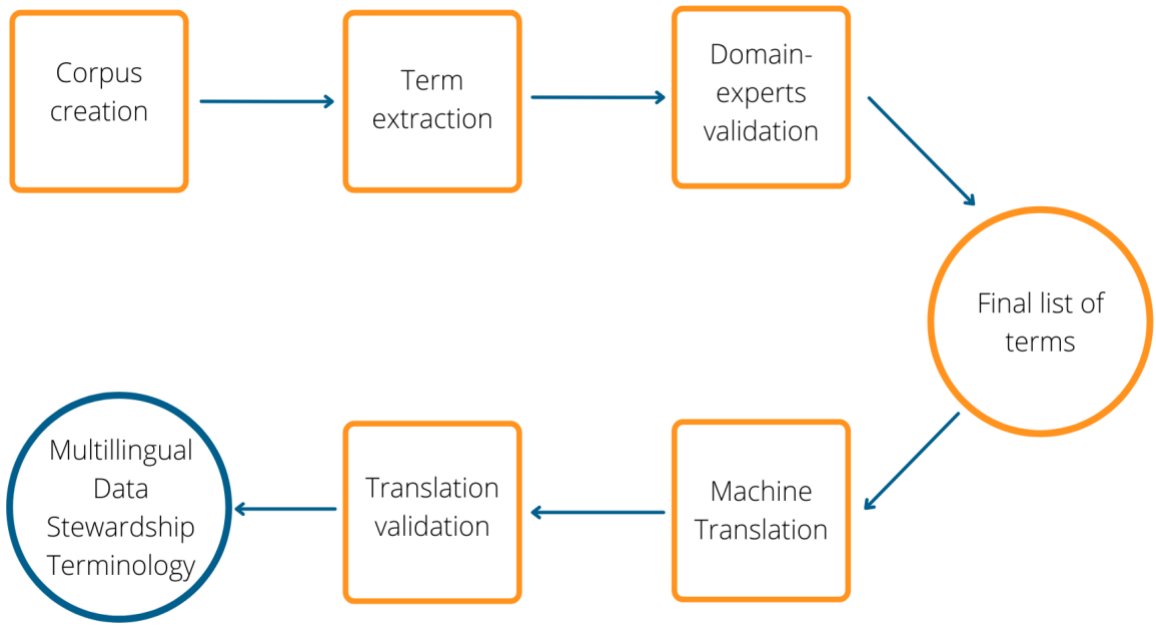


Figure 2. Workflow for the creation of the Data Stewardship Multilingual Terminology

3.2.1 Corpus creation

The first step towards the creation of the Data Stewardship terminology concerned the creation of a domain-specific corpus. To build the corpus, various documents pertaining to data stewardship and curation were collected. The corpus includes 70 documents among standards and recommendations for data stewardship and curation, deliverables and other technical documents. Open Access papers were selected. All the 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)[21] (15 Recommendations and 5 Supported Outputs) and through the OpenAIRE platform[22], by performing specific queries. All material that could be found and conformed to requirements (i.e., standards and technical documentation) was retained. No real selection was thus performed, as all the available documents found – if relevant – were included in the corpus, which reflects the status of available documents of September 2021. Since the chosen domain (Data Stewardship and Curation) 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 has been

[21] RDA Recommendations and Outputs catalogue: https://www.rd-alliance.org/recommendations-and-outputs/catalogue; [14 December 2021].
[22] OpenAIRE Searching Platform: https://explore.openaire.eu/; [14 December 2021].

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 the Appendix A.

3.2.2 Automatic term extraction

After building the domain-specific corpus, 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. 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. Four tools were identified that could serve the purpose: SketchEngine Keywords function (Kilgariff et al., 2014), TerMine (Frantzi et al., 2000), TermoStat (Drouin, 2003), TBXTools (Oliver and Vàzquez, 2015).

A brief description of the tools here follows:

- SketchEngine, Keywords function (Kilgariff et al., 2014): the Keywords function allows to automatically extract terminology from the corpus (focus corpus), with respect to a reference corpus; to this end, English Web 2020 (enTenTen20) was selected. Basic Sketch Engine settings were not modified, with the only exception of the parameter for minimum frequency, which was set to 3 in order to reduce the huge number of returned terms. The threshold of 3 is chosen as it represents an effective compromise between quantity and completeness: on the one hand, it allows to sufficiently reduce the number of candidate terms, while a threshold of 2 does not; yet it does not exclude too many of them, whereas if the threshold is set to 4 a significative number is neglected. Basic settings include case insensitivity, meaning that a lowercase version of the corpus is used, and the presence, in the extracted words, of only alphanumeric characters (words and letters). Extracted words are returned in two distinct categories: Keywords (single words) and Terms (multi-word expressions, MWEs).

- TerMine (Frantzi et al., 2000) Web Demonstration allows performing automatic term extraction on lightweight (max 2MB) raw text through the Web interface. If first PoS-tags the input text thanks to one of the two available PoS taggers: Tree Tagger version 3.1 (Schmid, 1994), most suited to generic texts, and GENIA Tagger version 2.1 (Tsuruoka et al., 2005), for texts from the biomedical sciences. TreeTagger was selected as the corpus at hand does not pertain to the biomedical domain. TerMine exploits the C-value domain-independent method for term recognition, which combines linguistic and statistical information (see Frantzi et al., 2000).

- TermoStat (Drouin, 2003) performs automatic term extraction 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 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. The four proposed different measures were tested (excluding raw frequency): log-likelihood, log-odds ratio, specificity (Lafon, 1980), chi-square. After experimenting with all the four measures, log likelihood was selected, although it does not substantially differ from the other statistical measures. No other parameters can be set (e.g., minimum frequency), so candidate terms occurring less than three times were manually excluded. TermoStat Web 3.0 was employed.

- TBXTools (Oliver and Vàzquez, 2015) is a Python class which performs terminology extraction based on either a statistical or a linguistic approach. Simple Python scripts calling the TBXTools class are provided.

    - *Statistical approach*
      Multiple parameters can be set. With respect to the size of n-grams to be extracted, it was decided to extract unigrams, bigrams, or trigrams. Stop-word filtering is performed thanks to a stop word list, already provided by TBXTools developers, which allows to eliminate all the candidates beginning or ending with a word from the list. Similarly, an inner stop words list eliminates candidates presenting one of the listed words. Another parameter allows to specify the minimum frequency of candidates to be extracted, in case of a large number of term candidates: it was set to 3. Some normalizations can then be performed. Case normalisation collapses the same term appearing with a different case into one same lowercase term. Nesting detection tries to spot shorter-term candidates that are not autonomous terms in and of themselves but are included in a longer term. Thanks to a rejection-list of regular expressions, listed patterns (mainly including non-word items) are excluded from the extracted terms. Candidates are then stored in a descending raw frequency order.

    - *Linguistic approach*
      Linguistic extraction 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. It was decided to load 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. The parameter specifying N-grams size was set as in the case of the statistical approach (unigrams, bigrams, or trigrams). Similarly, the parameter specifying the minimum frequency of candidates to be extracted, in case of a large number of term candidates, was set to 3. 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.

In order to select only the most suitable tools with respect to the task, an evaluation of the different tools was performed. However, as no gold standard was available for the corpus at hand, the calculation of recall represented an extremely time-consuming task. Precision could be calculated by manually checking the outputs of the tools (i.e., the lists of candidate terms). Yet, such a computation of precision could constitute a demanding task as well if performed for many different tools, as some of them clearly returned many non-relevant words at a first glance already. Consequently, a preliminary tool evaluation was designed that relied on an already annotated corpus, namely the ACTER dataset (Annotated Corpora for Term Extraction Research; Rigouts Terryn et al., 2020), version 1.4. The dataset, employed in TermEval2020, collects documents in 3 languages (English, French, Dutch) and is composed of 4 sub-corpora, concerning 4 different topics: corruption, equitation (dressage), heart failure, wind energy. For the purposes of the preliminary evaluation, the English wind energy subcorpus (58,654 tokens) was selected and considered as the gold standard of terms to be extracted from the list of terms marked as 'Specific Term' in the chosen ACTER subcorpus. Although this topic may seem distant from the one selected for this task, a good performance on this corpus can be expected to project also in other domain specific terminology extraction, including the one at hand.

Therefore, performances of the selected tools were evaluated with respect to such gold standard. A Python script allowed us to check if the extracted candidate terms occurred in the gold standard list of terms, thus carrying out a basic, yet consistent evaluation, in order to get a sense of the different features and key strengths of the tools. An overall and practical evaluation has been preferred, taking into account not only raw scores of precision and recall, but also the amount of extracted terms, since the following step consisted in a manual evaluation of the output. Table 6 presents the evaluation results.

	Extracted candidate terms	Correct candidate terms	Precision (%)	Retrieved candidate terms	Recall (%)	F-score (%)
TBXTools (statistical)	875	166	18,97	165	21,12	19,98
TBXTools (linguistic)	562	121	21,53	120	15,36	17,92
SketchEngine (single terms)	1670	75	4,49	75	9,60	6,11
SketchEngine (MWE)	444	116	26,12	116	14,85	18,93
TerMine	2778	329	11,84	329	42,12	18,48
TermoStat (specificity)	1392	228	16,37	226	28,93	20,91
TermoStat (chi-squared)	1391	228	16,39	226	28,93	20,92

TermoStat (log likelihood)	1391	228	16,39	226	28,93	20,92
TermoStat (log odds ratio)	1391	228	16,39	226	28,93	20,92
TermoStat (log likelihood) – minfreq=3	951	172	18,08	171	21,89	19,80

Table 6. Results of the preliminary evaluation on the wind subcorpus of ACTER dataset.

It can be observed from the table how the different statistical measures tested for TermoStat did not affect the results obtained, which were identical independently from the adopted measure. Therefore, log likelihood was selected as standard statistical measure when employing TermoStat, and its results with minimum frequency of terms set to 3 are presented as well, consistently with other tools. Overall, TermoStat proved to obtain good results, while extracting a number of terms which was not excessively high, thus being convenient for the needs of the task. TBXTools behaved in a similar way. The statistical approach proved slightly better than the linguistic one in terms of F-score, but it was decided to employ both for the extraction from the Data Stewardship corpus, as TBXTools resulted in contextually efficient and easy-to-use. With respect to SketchEngine, in the table it is distinguished between single terms and multi-word expressions. As for the former, an extremely low F-score correlated with a high number of extracted terms: poor results associated with an inconvenient number of terms to manually process. In case of multi-word expressions, results were better and aligned with the other tools. However, it was decided not to dissociate the two categories and thus not include SketchEngine in the subsequent work. As for TerMine, it was not possible to filter terms for raw frequency, as they are ordered according to C-value and the two scores do not correspond. However, by comparing the number of terms extracted by TerMine and the number of terms extracted by TermoStat before setting the parameter of raw frequency to 3 (see Table 6), it can be noticed how TerMine had a very good recall score, but returned too many candidate terms, thus making the manual revision for the purposes of precision a too demanding task.

Considering the results, TBXTools (both statistical and linguistic approach) and TermoStat were selected as the most suitable tools for the task. Clearly, 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.

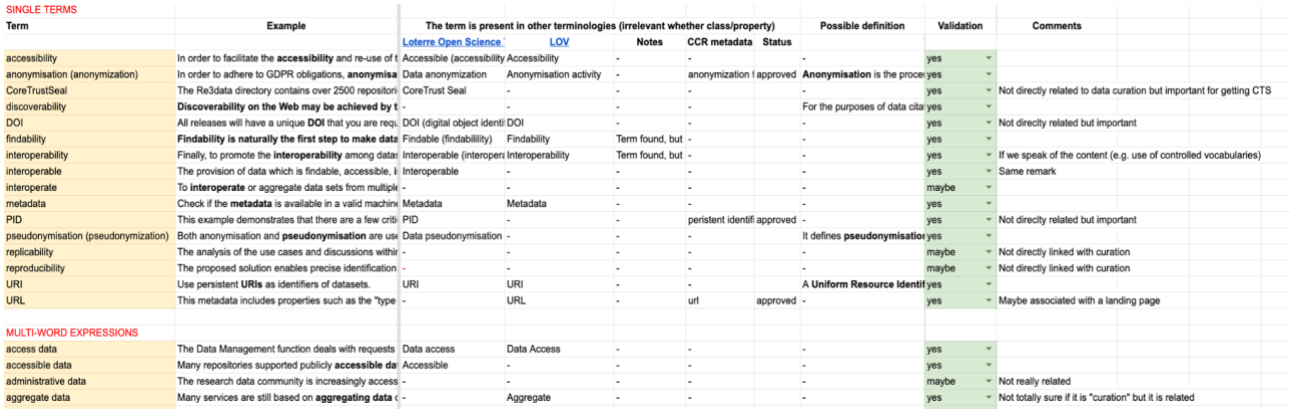
3.2.3 External validation

Afterwards the actual extraction on the Data Stewardship corpus was performed. The extracted candidate terms were manually revised so as to verify how many of them could be considered correct, by removing undoubted errors and non-terms. All these terms selected as correct, extracted by both TBXTools and TermoStat, have been combined in a single list of 277 candidate terms, which underwent an external validation by domain experts.

To this aim, a validation spreadsheet has been prepared, and structured as follows (Figure 3):

- candidate terms in the first column;
- an example of the term occurring in context;
- information about the presence of the term in other terminologies (Loterre Open Science Thesaurus[23], Linked Open Vocabularies[24], CLARIN Concept Registry[25]), that could be possibly checked. Mapping with terms4FAIRskills terminology had not been made at the time of validation, so it is not included in the validation spreadsheet;
- a possible definition, as extracted from the corpus – if found. Only few terms have one;
- validation column: three options to be selected (yes, maybe, no) answering the question "Is the term, as used in the example, a specific term of the domain of data stewardship?";
- a column for comments, if any (e.g., specify if the term is incomplete, if the form is not correct, link to other terminologies or possible definitions, etc.). If the option 'maybe' was selected, an explanation was requested.

The two domain experts, representing the validators, were provided with two distinct copies of the validation spreadsheet, to avoid any potential reciprocal conditioning.



SINGLE TERMS									
Term	**Example**	**The term is present in other terminologies (irrelevant whether class/property)**					**Possible definition**	**Validation**	**Comments**
		Loterre Open Science	LOV	Notes	CCR metadata	Status			
accessibility	In order to facilitate the **accessibility** and re-use of t	Accessible (accessibility	Accessibility	-	-		-	yes	
anonymisation (anonymization)	In order to adhere to GDPR obligations, **anonymisa**	Data anonymization	Anonymisation activity	-	anonymization (	approved	**Anonymisation** is the proce	yes	
CoreTrustSeal	The Re3data directory contains over 2500 repositori	CoreTrust Seal	-	-	-		-	yes	Not directly related to data curation but important for getting CTS
discoverability	**Discoverability on the Web may be achieved by t**	-	-	-	-		For the purposes of data cita	yes	
DOI	All releases will have a unique **DOI** that you are requ	DOI (digital object identi	DOI	-	-		-	yes	Not direcitly related but important
findability	**Findability is naturally the first step to make data**	Findable (findability)	Findability	Term found, but	-		-	yes	
interoperability	Finally, to promote the **interoperability** among data	Interoperable (interoper	Interoperability	Term found, but	-		-	yes	If we speak of the content (e.g. use of controlled vocabularies)
interoperable	The provision of data which is findable, accessible, i	Interoperable	-	-	-		-	yes	Same remark
interoperate	To **interoperate** or aggregate data sets from multipl	-	-	-	-		-	maybe	
metadata	Check if the **metadata** is available in a valid machin	Metadata	Metadata	-	-		-	yes	
PID	This example demonstrates that there are a few criti	PID	-	-	peristent identifi	approved	-	yes	Not direcitly related but important
pseudonymisation (pseudonymization)	Both anonymisation and **pseudonymisation** are us	Data pseudonymisation	-	-	-		It defines **pseudonymisatio**	yes	
replicability	The analysis of the use cases and discussions withir	-	-	-	-		-	maybe	Not directly linked with curation
reproducibility	The proposed solution enables precise identification	-	-	-	-		-	maybe	Not directly linked with curation
URI	Use persistent **URIs** as identifiers of datasets.	URI	URI	-	-		A **Uniform Resource Identif**	yes	
URL	This metadata includes properties such as the "type	-	URL	-	url	approved	-	yes	Maybe associated with a landing page
MULTI-WORD EXPRESSIONS									
access data	The Data Management function deals with requests	Data access	Data Access	-	-		-	yes	
accessible data	Many repositories supported publicly **accessible da**	Accessible	-	-	-		-	yes	
administrative data	The research data community is increasingly access	-	-	-	-		-	maybe	Not really related
aggregate data	Many services are still based on **aggregating data**	-	Aggregate	-	-		-	yes	Not totally sure if it is "curation" but it is related

Figure 3. Validation spreadsheet.

Subsequently adopting the following criterion compared the validation results. 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 answered '*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 and whether a proper definition could be found. For instance, Validator 1 validated the term big data, but this was not validated by Validator 2: since a

[23] Loterre Open Science Thesaurus: https://www.loterre.fr/skosmos/TSO/en/; [14 December 2021].
[24] Linked Open Vocabularies: https://lov.linkeddata.es/dataset/lov/; [14 December 2021].
[25] CLARIN Concept Registry: https://concepts.clarin.eu/ccr/browser/; [14 December 2021].

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. Sometimes validators' disagreeing answers were due to proposed possible definitions, as they remarked in the comments. However, as mentioned above, at this stage of the process only definitions found in the corpus were provided, although they were not always the most accurate to define the term and were later replaced by better ones. Therefore, in similar cases the pertinence of definitions was also considered when it was necessary to resolve disagreement. The final list of validated terms encompasses 260 entries.

In order to calculate the Inter-Annotator Agreement (IAA), linearly weighted Cohen's *k* statistical measure was selected, as it applies to cases where two raters are involved, is more reliable than mere percent agreement since it takes into account the agreement by chance, and allows to weight differently disagreement (e.g., '*yes*'-'*maybe*' are closer answers than '*yes*'-'*no*'). A0.08 Cohen's *k* was obtained, 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 proposed terminology resulting from the completion of this case study will 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 in order 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. The statistical approach of TBXTools returned 6582 candidate terms, resulting in a precision of 4.53%, a recall of 89.39% and an F-score of 8.661%. As for the linguistic approach, 3742 candidate terms were extracted: precision is thus 5.29%, recall is 73.48% and F-score is 9.87. Lastly, TermoStat obtains a precision of 8.50%, a recall of 83.71% and an F-score of 15.43%, as 3789 candidate terms were extracted. 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 lemmatization. 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.

3.2.4 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 210 distinct concepts.

Each concept was then provided with a definition. Definitions were 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. 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 (essentially in any consulted source: corpus, other terminologies, Web). 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, as anticipated while discussing validation. More specifically, its presence was checked in Loterre Open Science Thesaurus (developed at Inist-CNRS), in Linked Open Vocabularies (LOV) platform and in terms4FAIRskills (see Figure 4). ISO Online Browsing Platform (OBP)[26] allows for the querying of terms defined in ISO standardisation documents and was thus consulted as well, although not systematically; if a corresponding entry was found, it was linked with the term at hand.

[26] ISO Online Browsing Platform: https://www.iso.org/obp/ui/#search; [14 December 2021].

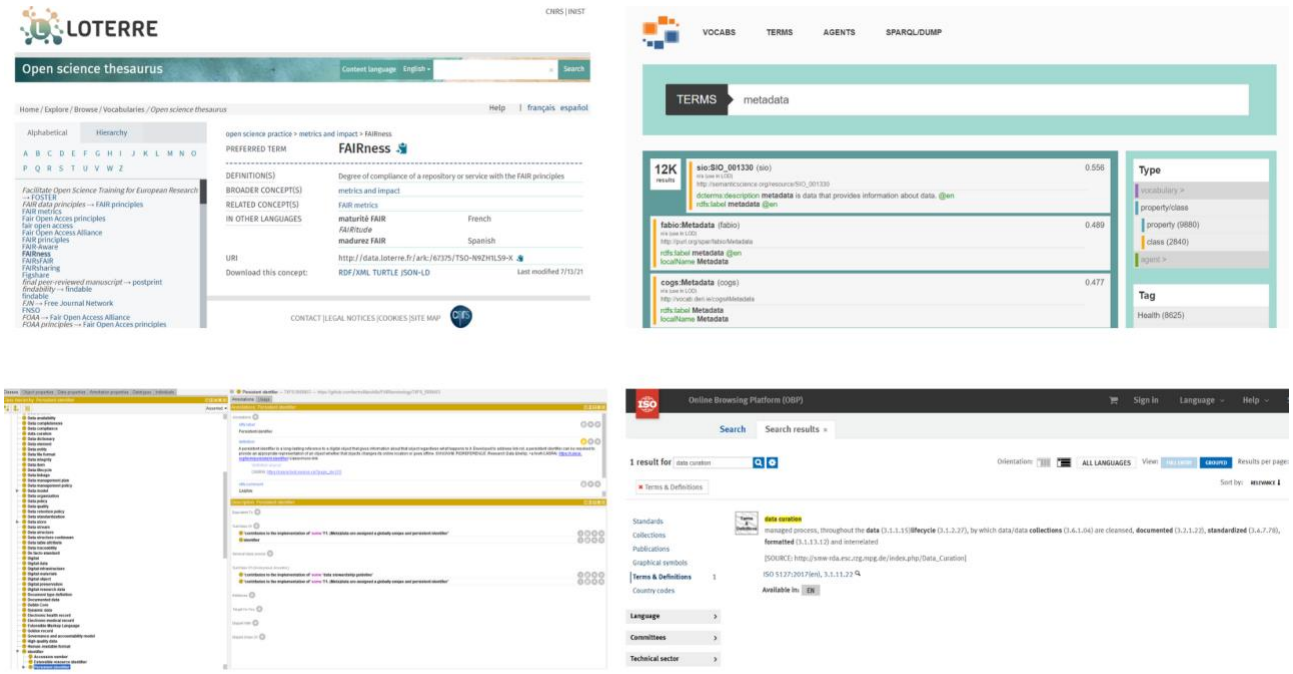


Figure 4. Loterre Open Science Thesaurus, Linked Open Vocabularies (LOV), terms4FAIRskills and ISO Online Browsing Platform (OBP) were consulted to link terms to existing terminologies.

As the intent was to create a multilingual terminology on Data Stewardship, it was then necessary to translate each pair of terms and definitions. Since Data Stewardship career paths and profiles are defined by the EOSC, they are always expressed in English. However, such concepts need to be introduced and adopted in the various European countries as well: for this reason, multilingualism is a key aspect in their dissemination, and translation of terms into different languages can play a key role in this respect. Conscious of this need to harmonise approaches throughout member countries, the EOSC defined a Task Force on Upskilling Countries[27] to engage in EOSC, in order to align Open Science education and skills initiatives across EOSC member countries. The adopted approach also allows us to investigate to what extent language technologies can accelerate the Upskilling Countries process, by focusing in particular on Machine Translation. It was decided to automatically translate the collected terms and definitions with a Machine Translation tool; to this end, Deep-L[28] was employed, since it resulted as the best performing MT tool among the ones tested in Topic 3.1.3. Given the similarities between the two topics, good performance was expected also for Topic 3.1.2. selected languages, that are the languages of the WP partners, are: Dutch, French, German, Greek, Italian, Slovenian (see Figure 5).

[27] EOSC Task Force on Upskilling Countries: https://www.eosc.eu/sites/default/files/tfcharters/eosca_tfupskillingcountriestoengageineosc_draftcharter_20210614.pdf; [14 December 2021].

[28] Deep-L: https://www.deepl.com/translator; [14 December 2021].

Concept ID		Term	Italian	French	German	Dutch
id_14	PrefLabel	data accessibility	accessibilità dei dati	accessibilité des données	zugänglichkeit der Daten	toegankelijkheid van gegevens
	AltLabel	accessibility	accessibilità	accessibilité	zugänglichkeit	toegankelijkheid
	Definition	FAIR principle encouraging to permanently stock data and metadata and simplify their access and/or download, by specifying access conditions (restricted or open access) and usage conditions (licence). A concrete way to address this principle is storing data in a deposit/warehouse ensuring their long term storage and making them accessible through a open protocol (e.g. HTTP) both by humans and machines.	Il principio FAIR incoraggia a conservare permanentemente i dati e i metadati e a semplificare il loro accesso e/o download, specificando le condizioni di accesso (accesso limitato o aperto) e le condizioni di utilizzo (licenza). Un modo concreto per affrontare questo principio è l'immagazzinamento dei dati in un deposito/magazzino che garantisca la loro conservazione a lungo termine e li renda accessibili attraverso un protocollo aperto (ad esempio HTTP) sia per gli esseri umani che per le macchine.	Le principe FAIR encourage le stockage permanent des données et des métadonnées et la simplification de leur accès et/ou de leur téléchargement, en spécifiant les conditions d'accès (accès restreint ou ouvert) et les conditions d'utilisation (licence). Une façon concrète de répondre à ce principe est de stocker les données dans un dépôt/entrepôt garantissant leur conservation à long terme et de les rendre accessibles par un protocole ouvert (par exemple HTTP) à la fois par les humains et les machines.	Der FAIR-Grundsatz regt dazu an, Daten und Metadaten dauerhaft aufzubewahren und ihren Zugang und/oder Download zu vereinfachen, indem die Zugangsbedingungen (eingeschränkter oder offener Zugang) und die Nutzungsbedingungen (Lizenz) festgelegt werden. Ein konkreter Weg zur Umsetzung dieses Grundsatzes ist die Speicherung von Daten in einem Depot/Lager, das ihre langfristige Speicherung gewährleistet und sie über ein offenes Protokoll (z. B. HTTP) sowohl für Menschen als auch für Maschinen zugänglich macht.	Het FAIR-beginsel moedigt aan om gegevens en metadata permanent slaan en de toegang en/of downloa ervan te vereenvoudigen door de toegangsvoorwaarden (beperkte of toegang) en de gebruiksvoorwaard (licentie) te specificeren. Een concr manier om dit beginsel toe te passe gegevens op te slaan in een depot/magazijn, zodat ze langdurig kunnen worden bewaard en ze toegankelijk zijn via een open proto (bv. HTTP), zowel voor mensen als machines.

Figure 5. The example of *data accessibility* illustrates how terms and definitions were automatically translated with DeepL into multiple languages (Greek and Slovenian are missing in the figure).

These translations, obtained automatically, underwent an external validation by native speakers. The intent was not to perform a proper evaluation of Deep-L, but to correct inaccurate translations and thus to generally assess how well the tool performed. 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 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.

Indeed, the whole 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.

Having said this, 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.

The data set is available at:
SSHOC Multilingual Data Stewardship Terminology -- http://hdl.handle.net/20.500.11752/ILC-567.

4. SKOSifying Results

The terminologies created within the context of Task 3.1 then needed to be represented and made available in SKOS, which is the recommended format discussed in task 3.5 "Data and Metadata Interoperability" in D3.1 "Report on SSHOC (meta)data interoperability problems" (Broeder et al., 2019), the discussions in the SSHOC Vocabulary Initiative and the underlying model of the SKOSMOS Vocabulary publication platform[29] -- as emerged from the analysis carried out in Task 3.1, Monachini et al. (2021) --, hosted by the OeAW partner.

Different solutions are used with respect to the operational SKOSification i.e. transforming the tabular vocabulary formats that are the outcome of translation and evaluation steps.

- For the metadata and data stewardship vocabularies, the CNR-ILC partner in collaboration with CNR-ISTI used a conversion tool. The SKOS-ifying mapping procedure for the flat table structure multilingual vocabularies that was used as a work-format has been implemented on the Jupyter Notebook platform and is available at: SKOS-ifying procedure -- http://hdl.handle.net/20.500.11752/ILC-566.

- CLARIN/WWI relies on the expertise of surveycodings.org, who will convert the excel-formatted occupational title vocabulary database into Extended Knowledge System (XKOS), an extension of the SKOS model for statistical classifications that is also compatible with the SSH Vocabulary Initiative recommendations. The XKOS description of the data of the occupational vocabulary will be available on the surveycodings website.

4.1 Implementing the SKOS mapping

The Data Stewardship terminology and the Multilingual metadata concepts were 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. The mapper is implemented in a Python Notebook and will be published in the SSH Open MarketPlace. The following tables (7 and 8) summarise the mapping rules for the transformation.

Data Stewardship terminology	SKOS Classes and properties
`Concept ID`	`skos:Concept`
`Term`	`skos:prefLabel @lang`
`Alternate Term`	`skos:altLabel @lang`

[29] Open source web-based SKOS browser and publishing software: http://www.skosmos.org [14 December 2021].

Term Definition	Skos:definition @lang
Source of Definition	dct:source
Loterre Open Science Thesaurus	skos:exactMatch
terms4FAIRskills	skos:note
Linked Open Vocabularies	skos:exactMatch
ISO	skos:note
Broader Concept	skos:broadMatch

Table 7. Mapping rules for the Multilingual Data Stewardship Terminology.

Metadata	**SKOS Classes and properties**
URI fragment	skos:Concept
Term	skos:prefLabel @lang
Term Definition	skos:definition @lang
Source	dct:source
URI	skos:exactMatch

Table 8. Mapping rules for Multilingual metadata.

4.2 SKOS Description of the resources

With respect to the Data Stewardship Multilingual 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 - DOI*), but it also provides a solution for handling alternative spelling variants (e.g. *anonymisation - anonymization*), often due to differences between UK and US English. It was discussed that for the purpose of this vocabulary there was no need to treat such variants as separate languages. 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 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 and has not been published in RDF yet. Therefore, when linking with one of these resources was possible, the `skos:note property` was used, which provides general documentation on a concept, and whose object is a literal and does not require a valid URI.

In line with the overall aim of the task, tIn line with the overall aim of the task, the general idea, in case of Data Stewardship Multilingual Terminology, was to test MT systems and provide translation and to test MT systems and provide translation and not to create a fully-fledged vocabulary with its own hierarchical structure. Thus, no internal hierarchy was defined, but a mild one was provided by linking the concepts extracted to broader terms in other terminologies, if possible.

An example of a concept (*anonymisation*) from the Data Stewardship Multilingual Terminology follows.

```
sshocterm:anonymisation_3 rdf:type skos:Concept;
      skos:inScheme sshocterm: ;
      skos:topConceptOf sshocterm: ;
      skos:prefLabel "anonymisation"@en;
      skos:altLabel "anonymization"@en ;
      skos:prefLabel "anonymisation"@fr ;
      skos:prefLabel "anonimisering "@nl ;
      skos:prefLabel "Anonymisierung"@de ;
      skos:prefLabel "ανωνυμοποίηση "@el ;
      skos:prefLabel "anonimizzazione"@it ;
      skos:prefLabel "anonimizacija"@sl ;
      skos:definition "Anonymisation is the process of removing any
information that could lead to the identification of the data
subject."@en;
      skos:definition "L'anonymisation est le processus qui consiste à
supprimer toute information pouvant conduire à l'identification du sujet
concerné (par ex. une personne, une institution, etc.)."@fr;
      skos:definition "Anonimisering is het proces waarbij alle informatie
die tot identificatie van de betrokkene kan leiden, wordt verwijderd."@nl;
      skos:definition "Unter Anonymisierung ist das Entfernen aller
Informationen zu verstehen, die zur Identifizierung des Datensubjekts
führen könnten."@de;
      skos:definition "Η ανωνυμοποίηση είναι η διαδικασία αφαίρεσης κάθε
πληροφορίας που θα μπορούσε να οδηγήσει στην ταυτοποίηση του υποκειμένου
των δεδομένων."@el;
```

```
      skos:definition "L'anonimizzazione è il processo di rimozione di
qualsiasi informazione che potrebbe portare all'identificazione
dell'interessato."@it;
      skos:definition "Anonimizacija je postopek odstranjevanja vseh
informacij, na podlagi katerih bi bilo mogoče identificirati posameznika,
na katerega se nanašajo osebni podatki."@sl;
      skos:broadMatch <http://data.loterre.fr/ark:/67375/TSO-MJLHMDNM-S>;
      skos:exactMatch <http://data.loterre.fr/ark:/67375/TSO-VNNFL8WB-H>;
      Skos:note "See also ISO/TS 17975:2015(en), 3.1" .
```

As regards the multilingual metadata, a similar approach was adopted. For each entry a `prefLabel`, a definition, and a source of the definition are specified. All `prefLabels` 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 CLARIN Concept Registry through the `skos:exactMatch` property.

4.3 Language and local specificities of the SKOS-ifying process

Some language-related and local specificities of the SKOS-ifying process of the Data Stewardship terminology, multilingual metadata and the occupation ontology were observed. For the Data Stewardship and metadata vocabularies it was clear that the subject-matter and languages involved did not warrant to make any difference between local language varieties, e.g. *anonymization* (en-US) and *anonymisation* (en-GB) at the level of needing to be presented as different languages. The en-US variant can be presented as an “alternative” label to the “preferred” en-GB spelled label, because such variants are relatively rare and the preferred label spelling is understandable and acceptable to all speakers. With respect to the occupational title vocabulary, there are different considerations. For the occupational titles as explained in section 3.1 it is essential that the occupational titles are specified per country which should also be each considered of equal value.

The SKOS standard and the foreseen vocabulary publishing platforms all support the IETF language tag formalism to which the multilingual ontology locale codes can be easily translated (eg. en_US maps to en-US).

5. Sustainability and Exploitation of Results

The lack of multilingual terminological resources and metadata in different domains constitutes an obstacle to the access and reuse of information. The added value for the end-user communities to provide multilingual metadata vocabularies and ontologies is to foster multilingual access to SSH content across different languages and improve discovery by non-native speakers.

Target users of the multilingual metadata, terminologies and ontologies produced in Task 3.1 are any SSH researchers and repositories needing to manage a wide range of SSH content. These resources are provided as freely and openly available data, findable through the VLO[30] and other CLARIN services and, hence, fully corresponding with the FAIR paradigm, in line with the EOSC principles and the wider policy context.

To overcome potential barriers to the exploitation, e.g., lack of update, maximise findability and hence usability and exploitation, CNR-ILC and CLARIN-IT provide the multilingual vocabularies in several formats, described with CMDI metadata through the ILC4CLARIN data-center, the national node of CLARIN-IT (the Italian CLARIN-B centre of the CLARIN-ERIC infrastructure). Additionally, the vocabularies are published in SKOS format via the SSH Vocabulary Commons publishing service[31] that permits browsing and accessing vocabularies by humans and API. Thus, ILC4CLARIN will be responsible for content and OEAW for the technical platform.

As concerns the multilingual occupation ontology, this will be hosted at surveycodings.org which offers a host of social science codings measuring individual and socio-economic variables. Also, the codings can be used as a resource for post-coding the open-ended questions during survey processing before its final release. In computer-assisted surveys the coding sets will be ready to use for respondent's self-selection, as well as interviewer's selection. Surveycodings ensures compliance to the FAIR principles promoted within EOSC. Links to generate the XKOS versions of the databases will be posted on the surveycodings website before the end of SSHOC.

Finally, long-term maintenance and updating represent a further challenge. In this respect, it will be crucial to monitor existing initiatives to encourage vocabulary reuse, and to promote community collaboration, thus avoiding duplication of efforts. Further testing of automatic extraction techniques and translation approaches will also be an important future direction.

[30] CLARIN VLO: https://www.clarin.eu/content/virtual-language-observatory-vlo; [14 December 2021].
[31] SSH Vocabulary Commons: http://vocabs.sshopencloud.eu/vocabularies [available from start 2022]

6. Conclusions

The deliverable presents three detailed case studies for each of the main topical areas of Task 3.1 aiming to investigate NLP and MT approaches in view of providing, translation of metadata concepts, multilingual vocabularies, terminology extraction across languages, multilingual databases.

Two of the topical case studies addressed here demonstrated that the contribution of NLP approaches and Machine Translation to the creation of multilingual resources is of critical importance (sections 2 and 3.2). 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. 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. Instead of translating from scratch, validators only need to verify the automatic translation, thus saving time and effort. The case study about providing a multilingual ontology with the correct male and female occupational titles, instead, showed that the automatic translation approach performs poorly (Section 3.1). None of the percentages of correct automatic translations of the English titles was above 20%.

From the three case studies a first set of recommendations can be derived 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 SSH Open Cloud after the lifetime of the project.

Table 9 summarises results and remarks as emerged from the selected case studies.

Topic	Question	Results observed	Resulting resource	Recommendations
Multilingual Metadata (Sect. 2)	Can MT tools offer an effective solution to translation tasks?	MT tools perform well, although their results need to undergo validation.	*Multilingual Metadata*: translated 232 CCR approved metadata concepts.	Promote community collaboration to encourage vocabulary reuse, avoid duplication of efforts and further test automatic translation approaches.
Multilingual Occupation Ontology (Sect. 3.1)	Can MT tools provide a solution to translation of job titles across languages,	Automatic translation approach performs poorly (correct automatic translations of the	*Multilingual Occupation Ontology*: the ontology consists of translated occupational titles	MT has proved to be a challenging task and should always be accompanied by a manual check from national experts, which

	specifically gender-specific titles?	English titles were under 20%). National labour market experts should check all MT translations.	of the master list (more than 4,200 titles).	requires serious consideration, planning and budget to complete.
Multilingual Data Stewardship Terminology (Sect. 3.2)	How can NLP techniques and MT approaches help create new multilingual terminological resources?	Automatic Term Extraction 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.	*Multilingual Data Stewardship Terminology*: 210 concepts about Data Stewardship, each with its definition, translated into multiple languages.	Promote community collaboration to encourage vocabulary reuse, avoid duplication of efforts and further test automatic extraction techniques and translation approaches.

Table 9. Overview of results.

As concerns validation, this appears to be an unavoidable step when exploiting MT tools, which definitely provide a solution to translation tasks but whose results need to be checked. Validation has to be performed by domain experts, also having knowledge of the topic besides being proficient in the language. Specifically for translating occupational titles, MT translations should all be checked by national labour market experts.

A few considerations are in order concerning the 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 CLARIN Concept Registry. In such contexts, the associated vocabularies usually already provide concept hierarchies and should be respected. Therefore, the Data Stewardship Multilingual Terminology is provided without its own hierarchy allowing more easy integration. Although, a partial hierarchy is obtained through linking it to other existing terminologies such as Loterre Open Science Thesaurus and terms4FAIRskills for convenience sake. 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 ask to be done[32]. In

[32] Terms4FAIRskills ontology URIs are currently only available via a GitHub dump, and thus not dereferencable; however the work on this ontology has been resumed under the direction of Laura Molly and Allyson Lister, and

the future, community collaboration with initiatives such as the EOSC Task Forces[33] will be promoted, to encourage vocabulary reuse, avoid duplication of efforts and further test the extraction and translation approaches adopted.

As to the format of the vocabularies -- that has been chosen following the recommendations emerged in this Task, cf. Monachini et al (2021) --, SKOS proved to have a sufficient degree of expressivity for what concerns the multilingual metadata concepts and the Data Stewardship Multilingual Terminology; more complex vocabularies are not needed to encode similar structures. The multilingual occupation ontology will be provided in XKOS, an extension of the SKOS model, still compliant to the recommendations emerged in this task.

Finally, the involvement in Task 3.1 allowed the partners to participate in discussions about vocabulary formats and publication platforms for optimal interoperability and management, thus advancing the awareness of FAIR principles for vocabularies and ontologies and the subsequent implementation of Task 3.1 vocabulary products. This provided an excellent opportunity for the involved organisations to make a step forward towards improved interoperability and sustainability with respect to vocabulary management.

the ontology is likely be published according to the Linked Open Data paradigms in early 2022 (personal communication).

[33] EOSC Task forces FAQ: https://www.eosc.eu/task-force-faq; [14 December 2021].

7. References

Artstein, R., and Poesio, M. (2008). Inter-coder Agreement for Computational Linguistics. Computational Linguistics, 34(4):555–596.

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., & Ďurčo, M. (2019). SSHOC D3.1 Report on SSHOC (meta)data interoperability problems (v1.0). Zenodo. https://doi.org/10.5281/zenodo.3569868.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. In: Terminology, vol. 9, no 1, p. 99-117.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic Recognition of Multi-word Terms. International Journal of Digital Libraries 3(2), pp.117-132.

ILO (2012). International Standard Classification of Occupations 2008, Volume 1 – Structure, Group Definitions and Correspondence Tables, International Labour Office (ILO), Geneva . Available at: https://www.ilo.org/public/english/bureau/stat/isco/isco08/ . (Accessed: 14 December 2021).

Johnson, M., Schuster, M., Le, Q., V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. Transactions of the Association for Computational Linguistics, 5, pp. 339-351. https://doi.org/10.1162/tacl_a_00065.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography, 1: 7-36.

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, http://hdl.handle.net/11234/1-2922.

Kulczycki, E., Guns, R., Pölönen, J., Engels, T. C. E., Rozkosz, E. A., Zuccala, A. A., Bruun, K. et al. (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): 1371–85. https://doi.org/10.1002/asi.24336.

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

Meng, C., Wessling, K., Mühleck, K., and Unger, M. (2020). EUROGRADUATE Pilot Survey Design and implementation of a pilot European graduate survey. Luxembourg: Publications Office of the European Union, 2020, doi:10.2766/629271.

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 (1.0.0). Zenodo. https://doi.org/10.5281/zenodo.4772741.

Monachini, M. (2020, November 25). SSHOC Considerations for the Vocabulary Platforms - Presentation of the MS08 results: Choice of Vocabulary publication platforms for SSHOC. Zenodo. https://doi.org/10.5281/zenodo.4290653.

Monachini, M., Jääskeläinen, T., Van Uytvanck, D., Van der Lek, I., Broeder, D., and Moranville, Y. (2021). MS08 Choice of Vocabulary Publication platform for SSHOC (1.0). Zenodo. https://doi.org/10.5281/zenodo.5181389.

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, Hissar, Bulgaria, pp. 473-479.

Padró, L., and Stanilovsky, E. (2012). Freeling 3.0: Towards Wider Multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012), 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 (LRE), Volume 54, Issue 2, pages 385-418. https://doi.org/10.1007/s10579-019-09453-9.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Tijdens, K. G. (2015). The design of a tool for the measurement of occupations in web surveys using a global index of occupations, Deliverable M21.4 of the InGRID project funded under the European Union's FP7 programme No: 312691, https://zenodo.org/record/1882080#.YZeLuNDMJPY.

Tijdens, K. G. (2019a). Database of occupational titles, with explanatory note. Deliverable 8.3 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. DOI: 10.13140/RG.2.2.22100.55687.

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., and Tsujii, J. (2005). Developing a Robust Part-of-Speech Tagger for Biomedical Text, Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS 3746, pp. 382-392.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144. arXiv.org e-Print archive. https://arxiv.org/abs/1609.08144.

List of Figures

List of Tables

Appendix A. Data Stewardship Corpus

List of documents included in the Data Stewardship corpus (see 3.2.1).

RDA Recommendations

1. Aryani, A. (2018). Data Description Registry Interoperability WG: Interlinking Method and Specification of Cross-Platform Discovery. https://doi.org/10.15497/RDA00003.
   Includes:
   a. Data Description Registry Interoperability WG. Interlinking Method and Specification of Cross-Platform Discovery – RDA Data Type Registries WG Output
   b. Data Description Registry Interoperability WG - Output Specification (supplementary material)
2. Burton, A., Fenner, M., Haak, W., and Manghi, P. (2017). Scholix Metadata Schema for Exchange of Scholarly Communication Links (Version v3). Zenodo. https://doi.org/10.5281/zenodo.1120265.
3. Burton, A., and Koers, H. (2016). ICSU--WDS & RDA Publishing Data Services WG Interoperability Framework Recommendations (1.0). https://doi.org/10.15497/RDA00002.
4. Caracciolo, C., Aubin, S., Jonquet, C., Amdouni, E., David, R., Garcia, L., Whitehead, B., Roussey, C., Stellato, A., and Villa, F. (2020). 39 Hints to Facilitate the Use of Semantics for Data on Agriculture and Nutrition. Data Science Journal, 19(1), 47. DOI: http://doi.org/10.5334/dsj-2020-047.
5. CoreTrustSeal (CTS). (2016). Core Trustworthy Data Repositories Requirements. DANS. https://doi.org/10.17026/dans-22n-gk35.
6. Dallmeier-Tiessen, S., Khodiyar, V., Murphy, F., Nurnberger, A., Raymond, L., Whyte, A., Bloom, T., Austin, C. C., Tedds, J., Stockhause, M., and Vardigan, M. (2016). RDA/WDS Publishing Data Workflows WG Recommendations (1.0). https://doi.org/10.15497/RDA00004.
   Includes:
   a. RDA/WDS Publishing Data Workflows WG - Connecting Data Publication to the Research Workflow: A Preliminary Analysis
   b. RDA/WDS Publishing Data Workflows WG - Final Outputs
   c. RDA/WDS Publishing Data Workflows WG - Workflows for Research Data Publishing: Models and Key Components
7. FAIR Data Maturity Model Working Group. (2020). FAIR Data Maturity Model. Specification and Guidelines (1.0). https://doi.org/10.15497/rda00050.
8. Hanahoe, H. (2020). Standards, Repositories and Policies: Facilitating Discovery, Adoption and Use. Executive summary of the RDA-adopted recommendations of the joint RDA/Force11 FAIRsharing WG.

9. Rauber, A., Asmi, A., van Uytvanck, D., and Proell, S. (2015). Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). https://doi.org/10.15497/RDA00016.
10. RDA Research Data Repository Interoperability WG. (2018). Research Data Repository Interoperability WG Final Recommendations. https://doi.org/10.15497/RDA00025.
11. RDA/TDWG Attribution Metadata Working Group. (2020). RDA/TDWG Attribution Metadata Working Group: Final Recommendations. https://doi.org/10.15497/rda00029.
12. Weigel, T., Almas, B., Baumgardt, F., Zastrov, T., Schwardmann, U., Hellström, M., Quinteros, J., and Fleischer, D. (2017). RDA Research Data Collections WG Recommendations (1.0). https://doi.org/10.15497/RDA00022.

RDA Supported Outputs

13. Andreassen, H. N., Berez-Kroeker, A. L., Collister, L., Conzett, P., Cox, C., Smedt, K. D., McDonnell, B., and Research Data Alliance Linguistic Data Interest Group. (2019). Tromsø Recommendations for Citation of Research Data in Linguistics (Version 1). Research Data Alliance. DOI: 10.15497/RDA00040.
14. Hrynaszkiewicz, I., Simons, N., Hussain, A., Grant, R., and Goudie, S. (2020). Developing a Research Data Policy Framework for All Journals and Publishers. Data Science Journal, 19(1), 5. DOI: http://doi.org/10.5334/dsj-2020-005.
15. Klump, J., Wyborn, L., Downs, R., Asmi, A., Wu, M., Ryder, G., and Martin, J. (2020). Principles and Best Practices in Data Versioning for All Data Sets Big and Small. Version 1.1. Research Data Alliance. DOI: 10.15497/RDA00042.
16. RDA-CODATA Legal Interoperability Interest Group. (2016). Legal Interoperability of Research Data: Principles and Implementation Guidelines. Zenodo. https://doi.org/10.5281/zenodo.162241.
17. Wittenburg, P., Hellström, M., Zwölf, C. M., Abroshan, H., Asmi, A., Di Bernardo, G., Couvreur, D., Gaizer, T., Holub, P., Hooft, R., Häggström, I., Kohler, M., Koureas, D., Kuchinke, W., Milanesi, L., Padfield, J., Rosato, A., Staiger, C., van Uytvanck, D., and Weigel, T. (2017). Persistent Identifiers: Consolidated Assertions. Status of November, 2017. Zenodo. https://doi.org/10.5281/zenodo.1116189.

Deliverables

18. Asmi, A., Cordewener, B., Goble, C. Castelli, D., Kühn, E., Pasian, F., Niccolucci, F., Glaves, H., Jeffery, K., Assante, M., Dovey, M., Manola, N., Juty, N., Blomberg, N., Jimenes, R., and Beckmann, V. (2017). D6.3: 1st Report on Data Interoperability: Findability and Interoperability (1.1). Zenodo. https://doi.org/10.5281/zenodo.3394145.

19. Bardi, A., Dimitropoulos, H., Foufoulas, Y., and Anagnostopoulou, A. (2021). D4.2 Transport Open Data: Properties and Specifications for Open Science. Zenodo. https://doi.org/10.5281/zenodo.4580461.
20. Baxter, R. (2019). EOSC-hub D2.8 First Data policy recommendations (Version 1). Zenodo. https://doi.org/10.5281/zenodo.4604670.
21. Behnke, C., Bonino, L., Coen, G., Le Franc, Y., Parland-von Essen, J., Riungu-Kalliosaari, L., and Staiger, C. (2020). D2.3 Set of FAIR Data Repositories Features (1.0 DRAFT). FAIRsFAIR. https://doi.org/10.5281/zenodo.3631528.
22. Budroni, P., Hönegger, L., and Bodlos, A. (2020). EOSC-Pillar D1.2 Data Management Plan. Zenodo. https://doi.org/10.5281/zenodo.3786326.
23. Davidson, J., Grootveld, M., Whyte, A., Herterich, P., Engelhardt, C., Stoy, L., and Proudman, V. (2020). D3.3 Policy Enhancement Recommendations (1.0 DRAFT). Zenodo. https://doi.org/10.5281/zenodo.3686901.
24. Devaraju, A., and Herterich, P. (2020). D4.1 Draft Recommendations on Requirements for Fair Datasets in Certified Repositories (1.0 DRAFT). Zenodo. https://doi.org/10.5281/zenodo.3678716.
25. Eurito. (2019). D3.1 Design of Data Collection Phase. Zenodo. https://doi.org/10.5281/zenodo.3243500.
26. Foggetti, N., Gerin Laslier, M., Di Giorgio, S., Haile Gebreyesus, N., Müller, S., van Nieuwerburgh, I., Romier, G., and Van Wezel, J. (2021). EOSC-Pillar D4.1 Legal and Policy Framework and Federation Blueprint. Zenodo. https://doi.org/10.5281/zenodo.4486610.
27. Gotz, A., Perrin, J.-F., Fangohr, H., Salvat, D., Gliksohn, F., Markvardsen, A., McBirnie, A., Gonzalez-Beltran, A., Taylor, J., and Matthews, B. (2020). PaNOSC FAIR Research Data Policy framework (1.1). Zenodo. https://doi.org/10.5281/zenodo.3862701.
28. Hugo, W., Le Franc, Y., Coen, G., Parland-von Essen, J., and Bonino, L. (2020). D2.5 FAIR Semantics Recommendations Second Iteration (1.0 DRAFT). Zenodo. https://doi.org/10.5281/zenodo.4314321.
29. L'Hours, H. (2018). Versioning Requirements for Curation and Access to New Forms of Data (01.00). Zenodo. https://doi.org/10.5281/zenodo.1406217.
30. L'Hours, H., Butt, S., Henriksen, G., Krejčí, J., Štebe, J., Myhren, M., Emery, T., and Bell, D. (2018). Generic High-level Workflows for the Curation of Different Forms of 'Big Data' (01.00). Zenodo. https://doi.org/10.5281/zenodo.1299685.
31. Molloy, L., Nordling, J., Grootveld, M., van Horik, R., Whyte, A., Davidson, J., Herterich, P., Martin, I., Méndez, E., Principe, P., Vieira, A., and Asmi, A. (2020). D3.4 Recommendations on Practice to Support FAIR Data Principles (1.1 DRAFT). Zenodo. https://doi.org/10.5281/zenodo.3924132.
32. Whyte, A., Engelhart, C., Bangert, D., Kayumbi-Kabeya, G., Lambert, S., Thorley, M., O'Connor, R., Herterich, P., and Davidson, J. (2019). D3.2 FAIR Data Practice Analysis (1.0 DRAFT). Zenodo. https://doi.org/10.5281/zenodo.3581353.

Other documents

33. Ashley, K. (2013). Data Quality and Curation. Data Science Journal. 12. GRDI65-GRDI68. 10.2481/dsj.GRDI-011.
34. Bailo, D., Paciello, R., Sbarra, M., Rabissoni, R., Vinciarelli, V., and Cocco, M. (2020). Perspectives on the Implementation of FAIR Principles in Solid Earth Research Infrastructures. Frontiers in Earth Science. 8. 10.3389/feart.2020.00003.
35. Baker, K.S., and Yarmey, L. (2009). Data Stewardship: Environmental Data Curation and a Web-of-Repositories. Int. J. Digit. Curation, 4, 12-27. https://doi.org/10.2218/ijdc.v4i2.90.
36. Bangert, D., Hermans, E., van Horik, R., de Jong, M., Koers, H., and Mokrane, M. (2019). Recommendations for Services in a FAIR data ecosystem. Zenodo. https://doi.org/10.5281/zenodo.3585742.
37. Borgman, C. L. (2012). Why Are the Attribution and Citation of Scientific Data Important? In: Uhlir, Paul and Cohen, Daniel (eds.). Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop. National Academy of Sciences' Board on Research Data and Information. National Academies Press: Washington DC. https://escholarship.org/uc/item/65b51130
38. Broeder, D., Budroni, P., Degl'Innocenti, E., Le Franc, Y., Hugo, W., Jeffery, K., Weiland, C., Wittenburg, P., and Zwolf, C. M. (2021). SEMAF: A Proposal for a Flexible Semantic Mapping Framework (1.0). Zenodo. https://doi.org/10.5281/zenodo.4651421.
39. Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., and Leadbetter, A. (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. International Journal of Digital Curation. 7. 10.2218/ijdc.v7i1.218.
40. Chue Hong, N., Cozzini, S., Genova, F., Hoffman-Sommer, M., Hooft, R., Lembinen, L., Marttila, J., and Teperek, M. (2020). Six Recommendations for Implementation of FAIR Practice (Final, pre-typesetting). Zenodo. https://doi.org/10.5281/zenodo.4065549.
41. Cichy, C. and Rass, S., "An Overview of Data Quality Frameworks," in IEEE Access, vol. 7, pp. 24634-24648, 2019, DOI: 10.1109/ACCESS.2019.2899751.
42. Devaraju, A., Mokrane, M., Cepinskas, L., Huber, R., Herterich, P., de Vries, J., Akerman, V., L'Hours, H., Davidson and J., Diepenbroek, M. (2021). From Conceptualization to Implementation: FAIR Assessment of Research Data Objects. Data Science Journal, 20(1), 4. DOI: http://doi.org/10.5334/dsj-2021-004.
43. Farias Lóscio, B., Burle, C., and Calegari, N. (2017). Data on the Web Best Practices. W3C Recommendation 31 January 2017. https://www.w3.org/TR/2017/REC-dwbp-20170131/.
44. Frosterus, M., Hansson, D., Dadvar, M., Kyriazis, I., Zapounidou, S., and Grant, F. (2021). Best Practices for Library Linked Open Data (LOD) Publication (Version 1). Zenodo. https://doi.org/10.5281/zenodo.4616759.
45. Gierasch, L. M., Davidson, N. O., Rye, K. A., and Burlingame, A. L. (2020). The Data Must Be Accessible to All. Mol Cell Proteomics. 2020 Apr;19(4):569-570. DOI: 10.1074/mcp.E120.001985.

46. Graber-Soudry, O., Minssen, T., Nilsson, D., Corrales, M., Wested, J., and Illien, B. (2021). Legal Interoperability and the FAIR Data Principles (1.0). Zenodo. https://doi.org/10.5281/zenodo.4471312.
47. Growth, P., Cousijn, H., Clark, T., and Goble, C. (2020). FAIR Data Reuse – the Path through Data Citation. Data Intelligence 2020; 2 (1-2): 78–86. DOI: https://doi.org/10.1162/dint_a_00030.
48. Hadziselimovic, E., Filip, D., and Lewis, D. (2021). Hybrid Framework for GDPR Data Compliance. https://doi.org/10.5281/zenodo.4922990.
49. Higgins, S. (2008). The DCC Curation Lifecycle Model. Int J Digit Curat. 3. 10.2218/ijdc.v3i1.48.
50. Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., Mietchen, D., Petrauskaité, R., and Wittenburg, P. (2018). Turning FAIR Data into Reality: Interim Report from the European Commission Expert Group on FAIR Data. https://doi.org/10.5281/zenodo.1285272.
51. Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., Stewart, C., Blake, M., Herndon, J., McGeary, T. M., and Hull, E. (2018). Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data. International Journal of Digital Curation, 13(1), 125-140. https://doi.org/10.2218/ijdc.v13i1.616.
52. Kalinauskaitė, D. (2017). To Be Findable, Accessible, Interoperable and Reusable: Language Data and Technology Infrastructure for Supporting the FAIR Data Approach. In CEUR Workshop proceedings: ICYRIME 2017: Proceedings of the Symposium for Young Researchers in Informatics, Mathematics and Engineering, Kaunas, Lithuania, April 28, 2017. Aachen: CEUR-WS, 2017, Vol. 1856, p. 21-26. https://hdl.handle.net/20.500.12259/35989.
53. Klump, J., Wyborn, L., Wu, M., Martin, J., Downs, R. R., and Asmi, A. (2021). Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles. Data Science Journal, 20(1), 12. DOI: http://doi.org/10.5334/dsj-2021-012.
54. Kowalczyk, S. (2018). Modelling the Research Data Lifecycle. International Journal of Digital Curation. 12. 331-361. 10.2218/ijdc.v12i2.429.
55. Laughton, P., and Plessis, T.D. (2013). Data Curation in the World Data System: Proposed Framework. Data Science Journal, 12, 56-70. DOI: 10.2481/dsj.13-029.
56. Lee, D. J., and Stvilia, B. (2017). Practices of Research Data Curation in Institutional Repositories: A Qualitative View from Repository Staff. PLoS ONE. 12. 10.1371/journal.pone.0173987.
57. Lenhardt, W. C., Ahalt, S., Blanton, B., Christopherson, L., and Idaszak, R. (2013). Data Management Lifecycle and Software Lifecycle Management in the Context of Conducting Science (Version 1). figshare. https://doi.org/10.6084/m9.figshare.791561.
58. Li, C., and Sugimoto, S. (2017). Provenance Description of Metadata Vocabularies for the Long-term Maintenance of Metadata. Journal of Data and Information Science, 2(2), pp. 41-55. https://doi.org/10.1515/jdis-2017-0007.
59. Padilla, T. (2016). Humanities Data in the Library: Integrity, Form, Access. D-Lib Magazine, 22. DOI: 10.1045/march2016-padilla.

60. Pandit, H. J., Debruyne, C., O'Sullivan, D., and Lewis, D. (2019). An Exploration of Data Interoperability for GDPR. International Journal of Standardisation Research, 16(1), 21. https://doi.org/10.5281/zenodo.3246453.
61. Pandit, H. J., O'Sullivan, D., and Lewis, D. (2018). GDPR Data Interoperability Model. 23rd EURAS Annual Standardisation Conference (EURAS). https://doi.org/10.5281/zenodo.3246439.
62. Pasquetto, I. V., Randles, B. M., and Borgman, C. L. (2017). On the Reuse of Scientific Data. Data Science Journal, 16, 8. DOI: http://doi.org/10.5334/dsj-2017-008.
63. Slouwerhof, I., Jetten, M., Elsenga, C., de Groot, N., Grootveld, M., Karvovskaya, L., Ras, M., de Smaele, M., Visscher, R., van den Berg, B., and Verheul, I. (2019). Dutch Data Curation Network. Report on the State of the Art of Data Curation in the Netherlands and the Feasiblity of Creating a Dedicated Dutch Data Curation Network /Een Nederlands netwerk voor datacuratie. Rapportage over de stand van zaken rond datacuratie in Nederland en de haalbaarheid van een speciaal datacuratienetwerk. (Final Version, December 2019). Zenodo. https://doi.org/10.5281/zenodo.3557237.
64. Thanos, C. (2017). Research Data Reusability: Conceptual Foundations, Barriers and Enabling Technologies. Publications. 5. 2. 10.3390/publications5010002.
65. Turp, C., Wilson, L., Pascoe, J., and Garnett, A. (2020). The Fast and the FRDR: Improving Metadata for Data Discovery in Canada. Publications, 8(2), 25. doi:10.3390/publications8020025.
66. Van Staden, S., and Mbale, J. (2012). The Information Systems Interoperability Maturity Model (ISIMM): Towards Standardizing Technical Interoperability and Assessment within Government. International Journal of Information Engineering and Electronic Business (IJIEEB). 4, pp. 36 – 41. 10.5815/ijieeb.2012.05.05.
67. Waithira, N., Mutinda, B., and Cheah, P.Y (2019). Data Management and Sharing Policy: the First Step Towards Promoting Data Sharing. BMC Medicine. 2019 Apr;17(1):80. DOI: 10.1186/s12916-019-1315-8.