CESSDA Work Plan 2020
Ontology Management System

D4b: ELSST Release Report

Document info

Dissemination Level	PU
Due Date of Deliverable	31/12/2020
Actual Submission Date	18/12/2020
Type	Report
Approval Status	Approved by CESSDA Tools & Services Working Group leader Mari Kleemola and CESSDA Technical Working Group leader John Shepherdson
Version	V2.0
Number of Pages	p.1 – p.11
DOI	10.5281/zenodo.4727781

Version history

Version	Date	Comment	Revised by
0.1	15.12.2020	First version created by UK Data Service, lead partner	
0.2	17.12.2020	Comments added by FSD and UKDS reviewers	UKDS
1.0	18.12.2020	First version submitted to CESSDA	
2.0	12.03.2021	Version 2 submitted to CESSDA following edits after WGL review	UKDS

Author List

Organisation	Name	Contact information
UK Data Service	Lorna Balkan	balka@essex.ac.uk
UK Data Service	Sharon Bolton	sharonb@essex.ac.uk
UK Data Service	Jeannine Beeken	jeannine.beeken@essex.ac.uk

Peer-review

Organisation	Name	Contact information
UK Data Service	Darren Bell	dbell@essex.ac.uk
Finnish Social Science Data Archive	Taina Jääskeläinen	taina.jaaskelainen@tuni.fi

Contents

Executive Summary

The European Language Social Science Thesaurus (ELSST) is a broad-based, multilingual thesaurus for the social sciences. It is owned and published by the Consortium of European Social Science Data Archives (CESSDA). During 2019/2020, work was undertaken to migrate ELSST from its previous bespoke UK Data Service (UKDS) thesaurus management system to the CESSDA Platform, using VocBench 3 as a back-end thesaurus management system and the Skosmos browser front-end system. This document describes the process undertaken to prepare ELSST content and documentation for the November 2020 release of ELSST.

Abbreviations and Acronyms

CC-BY-SA	Creative Commons Attribution-Share Alike licence
ELSST	European Language Social Science Thesaurus
ELSST TMS	ELSST Thesaurus Management System
HASSET	Humanities and Social Science Electronic Thesaurus
ICV	Integrity Constraint Validator
ISO	International Standards Organization
RDF	Resource Description Framework
SKOS	Simple Knowledge Organization System
SKOS-XL	Simple Knowledge Organization System eXtension for Labels
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
TMS	Thesaurus Management System
UKDS	UK Data Service
URI	Uniform Resource Identifiers
URL	Uniform Resource Locator
VOICE	Vocabularies in CESSDA (project)
XKOS	Extended Knowledge Organization System

1. About ELSST

The European Language Social Science Thesaurus (ELSST) is a broad-based, multilingual thesaurus for the social sciences.

ELSST is owned and published by the Consortium of European Social Science Data Archives (CESSDA). ELSST was originally based on the monolingual thesaurus, Humanities and Social Science Electronic Thesaurus (HASSET), of the UK Data Archive at the University of Essex. It has been further enhanced and extended through additional funding from the EU and the UK government. Since 2012, ELSST development has been funded by CESSDA through the CESSDA ELSST, Vocabularies in CESSDA (VOICE) and Metadata Office projects.

ELSST is used for data discovery within CESSDA and facilitates access to data resources across Europe, independent of domain, resource, language or vocabulary. The thesaurus covers the core social science disciplines: politics, sociology, economics, education, law, crime, demography, health, employment, information and communication technology and, increasingly, environmental science.

During 2020, the CESSDA Ontology Management System project oversaw the transfer of ELSST to CESSDA ownership and its migration from the previous UKDS software to the VocBench 3 thesaurus management system (based at UKDS) with a Skosmos browser front end (based at CESSDA). Both VocBench and Skosmos are open source tools. The process by which the successful release was facilitated and the relevant updates to ELSST documentation are described below.

2. Preparing and finalising ELSST content in VocBench 3

In the previous thesaurus management system (TMS), ELSST was stored as relational data in Microsoft SQL Server. A SKOS version of a section of the thesaurus was also available. Thesaural relationships defined in traditional rows and columns were converted into Resource Description Framework (RDF) 'triples'. RDF is used as a standard model for data interchange and interoperability across the internet. Data merging can be done using RDF features even if the underlying schemas are different. The framework uses URIs to name relationships between concepts as well as the two ends of the link (a 'triple'), forming a directed and labelled graph. [1]

VocBench 3 is based on RDF triples. While other import formats are available, it was decided to use the SKOS files generated by the ELSST TMS and import them into VocBench. First a concept scheme for ELSST in VocBench had to be drawn up. The concept scheme uses Simple Knowledge Organization Scheme (SKOS)[2] with Simple Knowledge Organization System eXtension for Labels (SKOS-XL)[3] for multilingual lexicalisations (Preferred Terms and

[1] https://www.w3.org/RDF/ (accessed 14 December 2020).
[2] https://www.w3.org/2004/02/skos/ (accessed 9 December 2020).
[3] https://www.w3.org/TR/skos-reference/skos-xl.html (accessed 9 December 2020).

Use For terms), and Extended Knowledge Organization System (XKOS)[4] for one of the note fields, i.e. the definition source. The XKOS field was adopted, due to the limited number of note fields available in SKOS. An equivalence between ELSST TMS elements and elements in the ELSST concept scheme in VocBench is shown in the Appendix.

ELSST TMS distinguishes between four different types of term status: 'Not Started', 'Untranslatable', 'Incomplete' and 'Complete'. 'Not Started' is where a Preferred Term of the source language, i.e. English, has not yet been translated; 'Untranslatable' is where no equivalent Preferred Term can be found in a target language'; 'Incomplete' is where the translation of the Preferred Term and its accompanying types of information is not yet finalised, and 'Complete' is where none of the other statuses apply. The first two statuses are marked syntactically in the value of the Preferred Term. 'Untranslated terms' are of the form XX_TERM, where XX is the 2-letter ISO 3166[5] country code of the target language, and TERM is the source language Preferred Term. For example, ES_ENERGY PRICES is the untranslated Spanish version of the source language term ENERGY PRICES. 'Untranslatable' terms are marked by adding '_UN' to the source language TERM in ELSST TMS, for example 'RIGHT OF WAY_UN'.

Both 'Not Started' and 'Untranslatable' terms were imported into VocBench in their current form. However, it was agreed that in future, there would be no need to mark either syntactically. 'Not Started' terms can be found via the VocBench Integrity Constraint Validator (ICV), and 'Untranslatable' terms will simply be marked as such in the language-specific Editorial Notes. The remaining term statuses, i.e. Both 'Incomplete' and 'Complete' terms were carried over into VocBench, without any formal distinction or marking. As with 'Untranslatable' terms, 'Incomplete' terms will be indicated via the 'Editorial Note'

In addition to mapping the data to the new concept scheme, a number of cleaning operations were performed on the data, including the following:

- HTML codes were found in the History notes and had to be removed
- Danish scope notes were language-tagged as 'dk' rather than 'da' and had to be corrected.

After uploading the data into VocBench, integrity checks were carried out on the data using SPARQL (SPARQL Protocol and RDF Query Language) queries in addition to checks using VocBench's built-in Integrity Constraint Validator (ICV). Checks included the following:

- Checks to ensure that all elements were in the correct SKOS fields according to the ELSST mapping drawn up
- Checks for polyhierarchies and how they were displayed
- Checks for broader and narrower term redundancy
- Ensuring that no content violated VocBench's Integrity Constraint Validator (ICV)

[4] https://ddialliance.org/Specification/RDF/XKOS (accessed 9 December 2020).
[5] https://www.iso.org/iso-3166-country-codes.html (accessed 15 December 2020).

- Ensuring that note fields were in the correct SKOS element type
- It was also decided to restart (date-stamped) the recording of the automated History or Change log

3. Transferring content from VocBench to Skosmos

The data were exported successfully from VocBench as RDF files (one for each ELSST language) and then assessed for readiness to import into Skosmos. At this point, some further modifications of the data were undertaken.

A mapping was made between the thesaurus elements used in VocBench and those used in Skosmos (see Appendix). It was decided to stick as closely as possible to the Skosmos standards, the main difference being the introduction of the 'Definition source'.

Some information that is available to ELSST internal developers and thesaurus managers in VocBench is not intended for publication in Skosmos and was deleted from files before import into Skosmos. This includes Editorial Notes, where translators can record any difficulties they encounter during the translation process, or where they can share draft translations with other translators.

The project team also decided that all Definition Sources[6] except those in the English version should not be uploaded to Skosmos, since they have not all been updated to conform with the new guidelines for the ELSST Bibliography of definition sources.[7]

It was also agreed to remove 'Not Started' Preferred Terms from Skosmos. Untranslated Preferred Terms are simply left blank. 'Untranslatable' terms are retained in Skosmos for the time being, but will be revised and replaced in future.

Once these operations were done, the final RDF files were delivered to the CESSDA main office technical team. Apart from adjusting the base URI as needed, the files were able to be loaded successfully into Skosmos, where they were marked as version 2020 (versioning/provenance).

4. Preparing and finalising ELSST documentation

In preparation for the release of ELSST two deliverable documents had to be finalised: *D3: ELSST Translation guidelines*, and *D3: ELSST User Guide.*

[6] Some ELSST concepts require further explanation/definitions to ensure clarity. Where these definitions are taken from published sources, we provide a three-letter acronym in the 'Definition Source' on the concept page and provide a key in the Bibliography of definition sources available on the ELSST website.
[7] https://elsst.cessda.eu/structure/bibliography (accessed 14 December 2020).

D3: ELSST Translation guidelines is designed for ELSST Translators. It forms part of the suite of training materials for ELSST Translators. Training is obligatory for ELSST translators in order to ensure translation quality. The guidelines are reviewed on a regular basis to make sure they are up to date.

The main modification to *ELSST Translation guidelines* was that the guidelines now use the Skosmos names for thesaurus elements instead of ELSST TMS names, as was previously the case. Since translators use VocBench to enter their translations, they also need to be familiar with VocBench names. The table of equivalence between all three systems (see the Appendix) was thus added to the guidelines.

The *ELSST Translation guidelines* also explain how translation status is now treated in ELSST (see Section 2 above).

D3: ELSST User Guide is part of the User Interface of the Skosmos implementation of ELSST, and is also available at https://elsst.cessda.eu/. Information that was still relevant was imported from the previous ELSST User Interface at the UKDS, including the Bibliography of Definition Sources. The main changes were the addition of a 'Concept scheme' section and updates to the 'Using ELSST' section. The Concept scheme section shows the equivalence between the ELSST concept scheme found in VocBench (and thus the ELSST SKOS files) and Skosmos elements, as well as the equivalences between these and the thesaurus elements in the ELSST TMS, for users who are familiar with the previous system. The 'Using ELSST' section contains Skosmos-specific information, instructions and examples of both the browse and search functionalities, which augments the basic information on search strategies found in the Skosmos 'Help' page. It also contains links to the API and ELSST download page, features that were not available in the previous system.

A new section entitled 'Release notes' was created to hold information not just about the latest content release, but previous ones as well. This information was found in different places in the previous ELSST User Interface (current changes under 'Changes' and previous changes in the CESSDA ELSST blog (https://www.ukdataservice.ac.uk/about-us/our-rd/cessda-elsst.aspx). The CESSDA ELSST blog will be discontinued, with future announcements on ELSST being made via the CESSDA portal.

5. Licensing ELSST

Previously, ELSST translators and those who wished to use ELSST in their systems had to contact the UKDS and sign a licence provided by the University of Essex. Use of ELSST has always been free of charge, but translations must be licensed in order to be included in the online system. As part of transfer of ownership to CESSDA, the licensing system had to be replaced. The organisations (CESSDA Service Providers) who had contributed a translation

for ELSST were notified that the University of Essex licence would be terminated. A new agreement was drawn up by CESSDA so that each translating organisation could license their translations to CESSDA instead prior to release. All licensees duly signed the agreement and so all 14 languages were included in the first release of ELSST to the CESSDA platform. ELSST is now covered by a widely-used standard, the Creative Commons Attribution-Share Alike licence (CC-BY-SA). Licensing information can be found on the 'Welcome' section of the CESSDA ELSST website.

6. Launch of ELSST

Once technical work and licensing were complete, documentation in place and all checks on the thesaurus undertaken, ELSST was launched on the current CESSDA platform on 16 November 2020. This version comprised the June 2020 content release, which had been the final content release in the previous system (the next content release is planned for Quarter 3, 2021). Full details of the content changes made for the June 2020 release are available in Metadata Office Task 2 2020 project deliverable *D1: Content report Notes for the ELSST release of 16 June 2020* [8] (as noted above, Metadata Office Task 2 covered content work in ELSST during 2020, while this project covered the technical work).

Since launch on the CESSDA platform, ELSST has now also been made available on the European Open Science Cloud (EOSC) Marketplace [9]. The Marketplace is a portal that provides access to many international resources from various research domains, including data analytics and metadata tools. The presence of ELSST there, alongside other CESSDA services such as the CESSDA Data Catalogue, will greatly increase its visibility to international social science researchers and data providers, and will raise its international profile and importance accordingly.

7. Summary

This document has described the background work undertaken to transfer content, ownership and licencing of ELSST from its previous incarnation to its launch as a CESSDA product. From the planning stage to release and beyond, the project team has worked in collaboration with CESSDA main office staff to ensure the robust construction and setup of technical systems, underlying data, and user interface to ensure a successful release of ELSST on the CESSDA platform. Within CESSDA, ELSST already facilitates access to data resources across Europe, independent of domain, resource, language or vocabulary. The

[8] Balkan, L. (2020) *D1: Content report: Notes for the ELSST release of 16 June 2020*, https://doi.org/10.5281/zenodo.4153696 (accessed 11 March 2021).
[9] https://marketplace.eosc-portal.eu/services/elsst-european-language-social-science-thesaurus (accessed 11 March 2021).

new opportunities for interoperability and linked open data (LOD) provided by VocBench and Skosmos tools will increase the popularity and uptake of ELSST even further.

References

Balkan, L. and Jääskeläinen, T. (2020) *D3: ELSST Translation guidelines*, report prepared for CESSDA.

Balkan, L. and Beeken, J (2020) *D3: ELSST User Guide*, report prepared for CESSDA.

Balkan, L. (2020) *D1: Content report: Notes for the ELSST release of 16 June 2020*, https://doi.org/10.5281/zenodo.4153696

Bolton, S., Balkan, L. and Beeken, J. (2020) *D1: ELSST Content Migration Report*, report prepared for CESSDA.

Appendix: Mapping between previous ELSST TMS elements and ELSST concept scheme in VocBench

VocBench	**Skosmos**	**Previous ELSST TMS**
skos:concept	Concept	Concept
skosxl:prefLabel	Preferred term	Preferred Term
skosxl:altLabel	Entry Term	Use For term
skos:broader	Broader concept	Broader term
skos:narrower	Narrower concept	Narrower term
skos:related	Related concept	Related term
skos:definition	Definition	Scope note
skos:scopeNote	Scope note	Use note
skos:editorialNote	Editorial note	Editorial note
skos:historyNote	History note	History note
xkos:additionalContentNote	Definition source	Scope note source