



Reusing lexical resources in the construction of a social science multilingual thesaurus

Chryssa Kappi ^a and Lorna Balkan ^b

^a Institute for Social Policy
National Centre for Social Research (EKKE)
Athens, Greece
ckappi@ekke.gr

^b UK Data Archive
University of Essex
Colchester, UK
balka@essex.ac.uk

Keywords: *Social sciences, lexical resources, multilingual thesauri*

Abstract

We describe how existing lexical resources were used in the development of the social science multilingual thesaurus European Language Social Science Thesaurus (ELSST) which was designed to be used for indexing and finding research data and metadata. The lexical resources of particular interest are *classifications, terminologies, thesauri, and controlled vocabularies*, collectively known as *Knowledge Organisation Systems (KOS)*. We first describe the properties of these different kinds of resources, which were designed for different purposes and audiences, and then discuss specific examples of classifications, terminologies and thesauri that were used in the development of ELSST. We propose possible solutions to problems we encountered that could enhance the reusability of these resources.

1 Introduction

Lexical resources are increasingly important tools for managing and retrieving data and information on the web. There is however a plethora of resources available, most of which have cost a lot of time and effort to produce. Much research effort, therefore, in recent years has gone into exploring ways in which to make these tools reusable and interoperable (van Hage et al., 2008; Lauser et al., 2008).

Here we describe how a variety of lexical resources that were designed for different purposes and audiences were used in the development of the social science multilingual thesaurus European Language Social Science Thesaurus (ELSST), and discuss some problems we encountered.



The lexical resources we looked at are knowledge organisation systems (KOS)¹. Zeng and Hodge (2011) define knowledge organisation systems as “*all types of schemes for organizing information and promoting knowledge management...Different families of KOS, including thesauri, classification schemes, subject heading systems and taxonomies, are widely recognised and applied in both modern and traditional information systems.*” In this paper we are concerned with KOS that are most relevant to archiving research data and metadata in the domain of social sciences, namely *classifications, terminologies* and *thesauri*. Most of these resources are controlled vocabularies, i.e. restricted sets of terms, as opposed to natural language, where there is no restriction on vocabulary use. Controlled vocabularies are intended to facilitate communication by resolving two problems associated with natural language, namely homography where two or more words have the same spelling, but mean different things, and synonymy, where two or more words have identical or similar meanings.

Lexical resources have always played an important role in data archiving and multidisciplinary and cross-national research in the social sciences. Prior to the construction of ELSST, European social science data archives had to rely on locally developed indexing tools including monolingual thesauri. The coverage of ELSST was designed to reflect the content of existing collections in European social science data archives, which means that aside from being multilingual, it was also desirable that it be compatible and interoperable with:

- a) international standard classifications
- b) social science documentation standards and relevant documentation software (e.g.: Data Documentation Initiative (DDI) (DDI Alliance, 2009))
- c) other related controlled vocabularies
- d) terminology in other related scientific fields (e.g. statistical terminology)

2 Short history of the construction of ELSST

ELSST has been developed over the last decade by the members of the Council of European Social Science Data Archives (CESSDA) to support access to its data catalogue. Data collections span sociological surveys, election studies, longitudinal studies, opinion polls, and census data, including international and European data such as the European Social Survey, the Eurobarometers, and the International Social Survey Programme. The development of ELSST was funded by a series of international projects² which aimed to create new or adapt existing controlled vocabularies for the representation of data and metadata in social sciences archives and other data organisations (Figure 1).

ELSST was originally developed from the English monolingual Humanities and Social Science Electronic Thesaurus (HASSET) created by the UK Data Archive at the University of Essex³. The process of adapting HASSET for ELSST was dictated by its multilingual character and its functionality in relation to comparative research, thus it involved stripping out lower level culture-specific terms and reworking term hierarchies, to make them more

¹ This term originated from the Networked Knowledge Organisation Systems Working Group at its initial meeting at the ACM Digital Libraries '98 Conference in Pittsburgh, Pennsylvania according to Hodge (Hodge, 2000).

² For example the LIMBER, MADIERA and CESSDA PPP projects (Miller & Matthews, 2001; Alvheim, 2006; CESSDA PPP, 2008-2010).

³ Currently ELSST is being used for searching CESSDA collections; It is available for the general public to view at the following web page: <http://elsst.esds.ac.uk/login.aspx>. Anyone wishing more information on ELSST should contact Lucy Bell: ljbell@essex.ac.uk

relevant across cultures. In the latter phase of its construction, as emerging CESSDA and other European projects were increasingly aiming at standardisation of research and archiving methods in comparative research, the need to make ELSST interoperable with co-existing relevant lexical resources led to its being reviewed and restructured (CESSDA PPP, 2008-2010). All work, including the search for other language equivalents, was undertaken by a multilingual, multidisciplinary team, consisting of language experts and subject specialists.

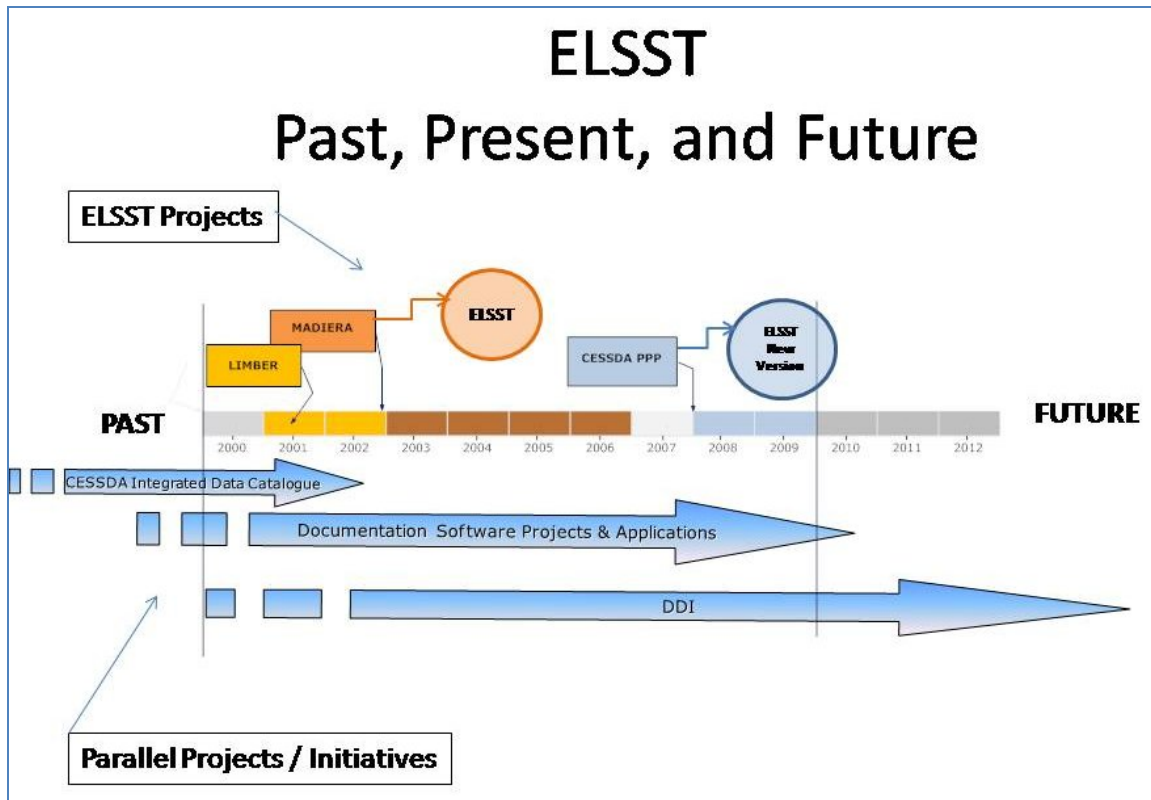


Figure 1. ELSST timeline (this figure is an adaptation from Miller (2004))

ELSST is currently available in nine languages – Danish, English, Finnish, French, German, Greek, Norwegian, Spanish and Swedish, with over 3,000 terms for the majority of its languages. Users can browse the thesaurus, perform free-text or keyword searches in their preferred language and retrieve all relevant resources not only in that language but in any of the nine supported languages. In this way, ELSST promotes cross-national comparative research in the social sciences.

3 Overview of lexical resources

In this paper, we focus on knowledge organisation systems (KOS), which are key resources for information retrieval. There is no one definition of KOS, although it is generally agreed that they include at least the following subtypes, as defined by Hodge “*term lists, which emphasize lists of terms often with definitions; classifications and categories, which emphasize the creation of subject sets; and relationship lists, which emphasize the connections between terms and concepts.*” We show this schematically in Figure 2, where we have added terminologies to Hodge’s category ‘term lists’.

We have also added controlled vocabularies to Hodge’s classification. Thesauri are a type of controlled vocabulary. Controlled vocabularies resolve ambiguity by adding qualifiers to

homographs. For example in ELSST, the medical sense of ‘labour’ is distinguished from the work-related sense by the addition of a qualifier in brackets: LABOUR (PREGNANCY) and LABOUR (WORK) respectively. Synonymy in controlled vocabularies is handled by identifying one term as the preferred or controlled term and listing terms with identical or similar meanings as synonyms.

Figure 2 is intended to show that controlled vocabularies can span each category of KOS.

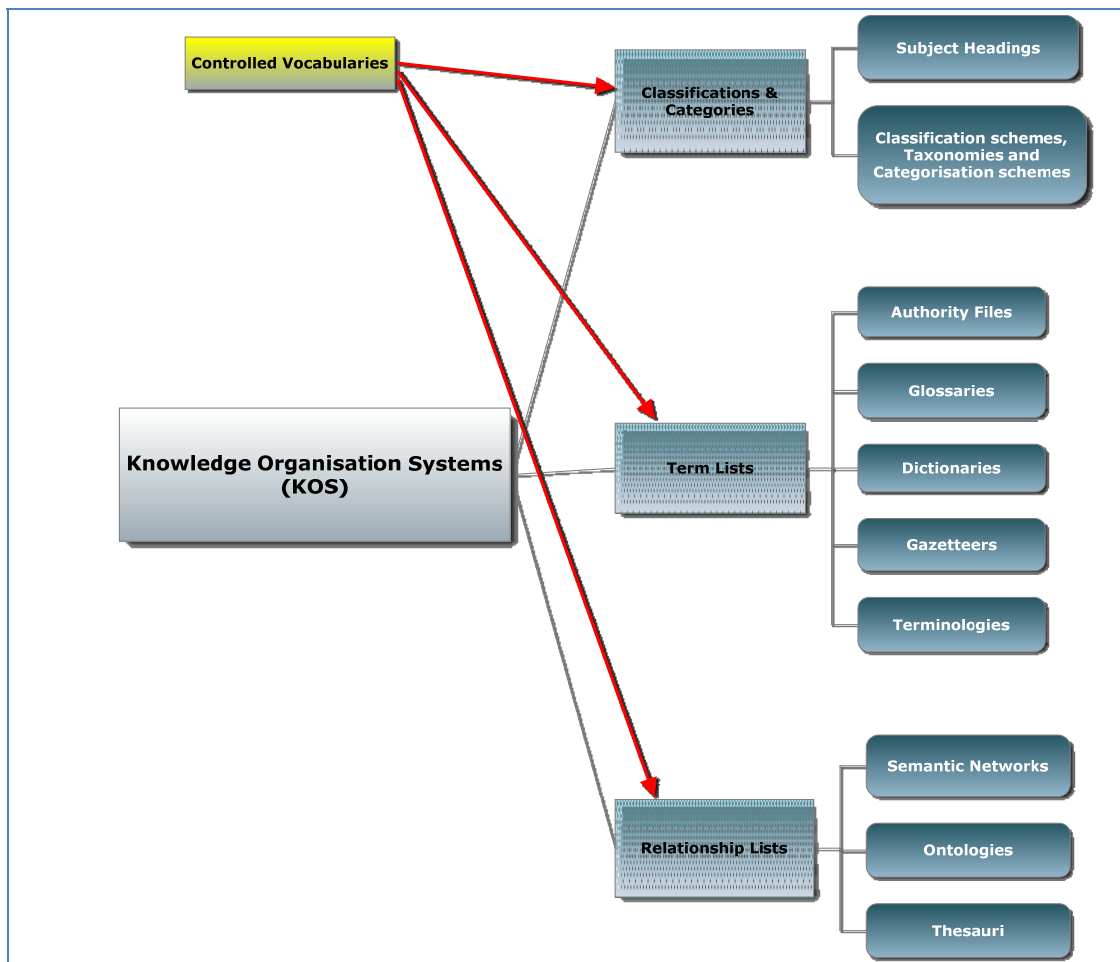


Figure 2. A classification of Knowledge Organisation Systems (KOS)

3.1 Lexical resources: definitions

3.1.1 Classifications

Classification is the act of any grouping of items under specific organising principles or criteria (classifiers). It is defined by Hayek (Hayek, 1976) as a process “*in which on each occasion on which a certain recurring event happens it produces the same specific effect, and where the effects produced by any one kind of such events may be either the same or different from those which any other kind of event produces in a similar manner [...]*”. Classifications are the product of the classification process and can be described as a set of terms representing concepts or classes, listed hierarchically or in some other systematic fashion in order to display relationships among them.

Classifications are the oldest type of KOS, in the sense that they predate the internet; they can be traced back to the first attempts at organising scholarly documents, serving the



communication needs of the knowledge community –“*the Republic of Science*”, as Guédon calls it (Guédon, 2008). This led to the production of the early classification systems, such as the *Dewey Decimal Classification* (DDC) (DDC, 1989), first published in 1876, and the *Library of Congress Classification* (LCC) (The Library of Congress, 2009). Early classification systems served to promote communication between two parties: indexer and information retriever. For early 20th century libraries this was a rational model, because the librarian was meant to serve the academics and researchers across all scientific disciplines without being an active agent of the research process. Nowadays the indexer tasked with classifying objects faces greater challenges, including the vast amount of information available, its exponential rate of growth, and the trend towards ever greater specialisation and integration of knowledge made possible by technology (Kappi, 2009). Problems are compounded by the fact that the information retriever is often not a trained librarian, but more often a primary researcher, unlike the early 20th century model.

3.1.2 Terminologies

ISO 1087 (ISO, 1990) defines terminology as a “*set of designations belonging to one special language*”, where designations can be terms, appellations or symbols. A terminology usually includes definitions of the designated objects (i.e.: concepts) and possibly other information such as sources and examples. While in the past many terminologies were word-based, the more common method nowadays is to start from the concept.

3.1.3 Thesauri

A thesaurus consists of a structured set of terms. Terms can be either descriptors, i.e. controlled terms that are used for indexing purposes, or non-descriptors (or synonyms), which cannot be used for indexing purposes but which act as pointers to the descriptors. Descriptors denote one and only one concept. Terms are related via a fixed set of relationships which typically include hierarchy (represented by the notation BT (broader term) and NT (narrower term)), synonymy (UF (Used for)) and association or relatedness (RT (related term)). Scope notes (SN) are also provided if necessary to clarify meanings or usage of a term. International standards exist for thesaurus construction (see, for example (ISO, 1985; ISO, 1986)).

Characteristics	Lexical Resources			
	Classifications	Thesauri	Terminologies	Controlled Vocabularies
Definitions	Optional	Optional	Optional	Optional
Relations Schema	Optional	Necessary	Optional	Optional
Domain Specificity	Optional	Optional	Necessary	Optional
Function	Categorizing objects with common characteristics	Organising terms and defining their relationships for indexing and information retrieval	Promoting communication in specific knowledge domains	Controlling use of terms in disparate information sources

Table 1. Lexical resources: a comparison



Table 1 above presents a comparison of lexical resources based on a set of key characteristics. There is, however, no hard and fast distinction between the different types of lexical resources. Thus, for example, classifications may simply provide a structured view of a set of objects without any definition of concepts and relations; or they may take the form of thesauri where concepts and relations are defined. Likewise, terminologies with hierarchical structures can be considered highly specialised classifications; and a controlled vocabulary “[...] *can range from a short list of clearly defined, mutually exclusive, and exhaustive terms, which are the only choices for usage in a specific context [...] through a classification to something as complex as a thesaurus with thousands of terms and term relationships.*” (Granda, Kramer & Linnerud, 2008).

3.2 Lexical resources: functions

Table 1 shows the main functions of each of the above lexical resources. However, the different types of resources may be regarded as complementary to each other, serving different purposes. For example, in medical practice, clinical terminologies are used for specific and detailed annotations of cases, while classifications may be used for organisational and administrative record-keeping for the care provided (Campbell & Giannangelo, 2007). Difference in function of lexical resources is reflected in differences in their content and structure. For example, terminologies tend to be much more fine-grained than classification systems, since the whole point of the latter is to provide broader ways of classifying objects.

All the above lexical resources help to promote consistency in language use within (and possibly beyond) the organisation in which they were created, and thus improve communication. Increasingly, as they become more amenable to machine interpretation they are also finding new uses in computer applications such as text mining. Controlled vocabularies in particular are set to play a critical role in the semantic web.

Multilingual lexical resources have an additional role as support to translators, and as tools to allow access to data across geographical or cultural divides. As the Semantic Web grows, so too does the demand for multilingual accessing and querying of knowledge repositories and linked data that are becoming available.

There is also an increasing trend towards standardisation of research data and metadata (including controlled vocabularies) worldwide. Examples include the DDI (DDI Alliance, 2009) an international alliance which works on a number of relevant projects, and the existence of international standard classification systems, such as International Standard Classifications (ISC) for specific subject domains, e.g.: International Standard Classification of Occupations (ISCO) (ILO, 2011), International Classification of Functioning, Disability and Health (ICF) (WHO, 2010b), International Classification of Diseases (ICD) (WHO, 2010a), International Standard Classification of Education (ISCED) (UNESCO, 2006), etc.

4 Concept analysis as foundation for building lexical resources

4.1 Overview

Common to most of the lexical resources discussed in Section 3 above, is the need for conceptual analysis to achieve high quality. By conceptual analysis we mean the breaking down of the concepts represented by the linguistic labels or terms into their component parts to better understand and/or define them.

Concept analysis is particularly important for establishing multilingual equivalence between terms, since the meaning of concepts may be culture-dependent.



While concept analysis is acknowledged as necessary for high quality lexical resources, there is no agreed methodology for how to do it. All approaches to concept analysis, however, start with three steps, regardless of the knowledge domain: (1) define the goal(s), (2) define the scope, i.e. the domain boundaries and (3) list the concepts to be analysed (see Nuopponen, 2010a; Nuopponen, 2010b; Nuopponen, 2011). The essential aim of concept analysis, irrespective of motivation and method, is the *transfer of meaning*. This aim is of particular importance in the social sciences domain since meaning is an essential element in any communication act (see for example Chouliaraki & Fairclough, 1999; Harbsmeier, 2007). Guarino (1995) argues that to accomplish the transfer of meaning, it is not enough to identify concepts and their designation -the relationships between concepts must also be defined. Context analysis also involves examining the context in which concepts are used (e.g. sources documents) in order to identify the key characteristics which help to define and differentiate them. The process requires the skills of both linguists and domain specialists (see for example ISO, 2009; Priss, 1996; Priss, 2006).

4.2 Concept analysis for lexical resources in the social sciences

Social science terms have a number of characteristics that make them ideal candidates for concept analysis. Firstly, many social science terms are intrinsically vague or ambiguous, even at a monolingual level. This is because there is a large overlap with general language.

Secondly, there is a tendency, particularly in applied social sciences, to treat terms as though they express invariant concepts which can be used and re-used for research purposes irrespectively of temporal, spatial, social or other context. However, the meaning of social science terms, like any other words of a language, can evolve over time, as a result of societal changes. For example, what 'old age' denotes today is not the same as it denoted 20 years ago, and what it possibly might denote in another 20 years, as life expectancy rises.

This creates problems for the design, analysis, preservation and dissemination of research data. In field research, for example, where concept definitions do not form part of the documentation, variables and units of measurement are assigned linguistic labels, i.e. terms, which 'imply' the meaning of underlying concepts, without defining them. Examples include 'childhood', 'household' and 'special needs'. Thus, ideally, any type of social study should be accompanied by a formal concept analysis process for its study objects. In practice, concepts are only defined where required by the study, as in the case of datasets earmarked for comparative research.

4.3 Concept analysis in ELSST

In ELSST, the English language terms were used as the source for all other language equivalents. This was not to impose dominance of English over any of the other languages, but to try to ensure that a designation for the same concept was sought in every target language. We refer to 'target language equivalents' rather than translations, since finding target language equivalents was viewed as being more akin to a mapping process than a translation process (see Doerr, 2001). Functional equivalence, rather than strict equivalence was aimed for. Equivalent terms are expected to allow retrieval of data on the same topic across archives and languages. In practice this means that target language terms are acceptable that express partial as well as exact equivalence with the source term. (Partial equivalence in ELSST is defined as cases where source language and target language terms are generally regarded as referring to the same concept, but where one of the terms strictly



denotes a slightly broader or narrower concept. More information on different types of term equivalence in ELSST can be found in Balkan et.al., (2010)). Another consequence of this strategy is that all language versions of ELSST strive to be ‘equally authentic’, in the sense adopted by European Union language policy (see Athanasiou, 2006; Fidrmuc et al., 2007; Luttermann, 2011), i.e. treated as originals and not as translations.

In ELSST concept analysis was carried out where difficulties arose in finding target language equivalents. Except in straightforward cases it was not possible to find a target language equivalent for a term without a source term definition or scope note. To find a definition, the source language team had to examine synonyms and related terms, and, where necessary, review the research material that had been indexed with the term in order to pin down its exact meaning. The term, together with its scope note, was then passed to the target language teams to find equivalent terms. They in turn often had to perform concept analysis on candidate target language terms in order to establish the degree of equivalence with the source term. Cross-team collaboration was frequently required to arrive at a solution. All terms and their scope notes were also peer-reviewed by field experts.

Another case where concept analysis was required in ELSST was where efforts were made to align ELSST with existing lexical resources (see Section 5 below).

Developing multilingual lexical resources has much in common with establishing interoperability between monolingual resources. Tudhope, Koch, & Heery (2006) note that “*Mapping is a key requirement for semantic interoperability in heterogeneous environments. Although schemas, frameworks and tools can help, detailed mapping work at the concept level is necessary, requiring a combination of intellectual work and automated assistance. Significant effort is required for useful results.*” The examples described below were based on intellectual work alone. As with Tudhope et al., mapping between terms was performed at the concept level.

5 Multilingual lexical resources and ELSST

5.1 Classifications

International classifications were consulted in the development of ELSST particularly for cases where there was a wide discrepancy in terms due to differences in institutional structures and systems. An example is the domain of education, where the International Standard Classification of Education (ISCED-97) (UNESCO, 2006) was consulted.

5.1.1 The International Standard Classification of Education (ISCED-97)

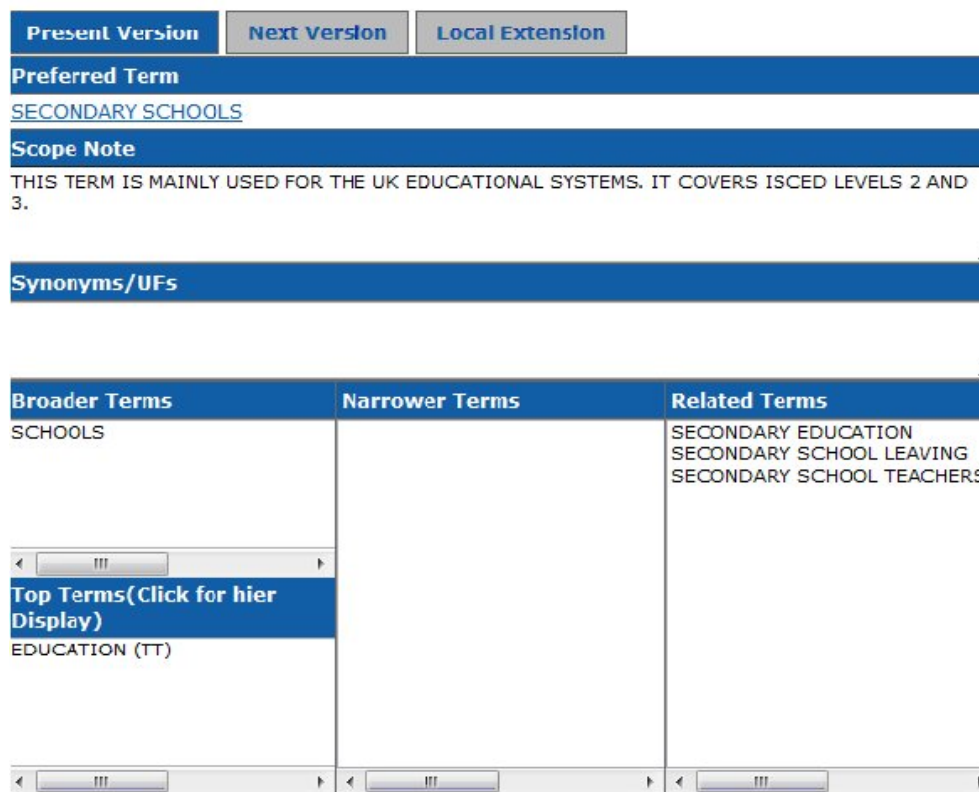
International Standard Classifications (ISC), produced by United Nations Educational, Scientific and Cultural Organization (UNESCO), serve as tools for comparisons and are considered to be “[...] *the foundations on which all statistical systems, national as well as international, are built*” (Eurostat, 1999:5). It is intended that comparability in ISC be achieved not only via their structure, but also by means of their conceptual transparency.

ISCED was specifically designed “[...] *to greatly improve the comparability of education statistics – as data collected under this framework will allow for the comparison of educational programmes with similar levels of educational content – and to better reflect complex educational pathways in the OECD indicators*” (OECD, 1999:3). ISCED-97 is used in cross-national surveys for the purposes of statistical reporting, creation of research tools, comparisons, policy making, etc. It has been extensively tested for its cross-national

applicability (see for example Eurostat, 1999; OECD, 1999; Schneider, 2008). National statistical institutes (NSI) provide resources on educational subjects in compliance with the ISCED-97 levels of education (Eurydice, 2012).

Mapping the ELSST EDUCATION hierarchy, which includes EDUCATIONAL SYSTEMS and EDUCATIONAL INSTITUTIONS, to ISCED-97 demanded long discussions among source and target language teams about educational concepts and their definitions. Specialised resources, such as the European Glossary on Education were also consulted. In view of the difficulties in mapping ELSST to ISCED-97, restructuring ELSST according to ISCED-97 levels was discussed but rejected, on the grounds that little would be gained.

Most problems arose from the fact that ISCED-97 refers to systems that are too culture-specific. In ELSST it was not always possible to find equivalent terms in one or more of the target languages since the corresponding concept was missing. Several solutions were adopted in such cases, such as adding metadata support (i.e.: definitions and scope notes) or employing descriptive phrases or neologisms⁴. For example, consider the term SECONDARY SCHOOLS and its scope note in ELSST (Figure 3).



The screenshot shows a web-based interface for the ELSST term 'SECONDARY SCHOOLS'. At the top, there are three tabs: 'Present Version' (selected), 'Next Version', and 'Local Extension'. Below the tabs, the 'Preferred Term' is 'SECONDARY SCHOOLS'. The 'Scope Note' states: 'THIS TERM IS MAINLY USED FOR THE UK EDUCATIONAL SYSTEMS. IT COVERS ISCED LEVELS 2 AND 3.' Below this is a section for 'Synonyms/UFs'. At the bottom, there is a table with three columns: 'Broader Terms', 'Narrower Terms', and 'Related Terms'. The 'Broader Terms' column contains 'SCHOOLS' and 'EDUCATION (TT)'. The 'Related Terms' column contains 'SECONDARY EDUCATION', 'SECONDARY SCHOOL LEAVING', and 'SECONDARY SCHOOL TEACHERS'. There are also navigation arrows and a 'Top Terms (Click for hier Display)' button.

Figure 3: ELSST term ‘SECONDARY SCHOOLS’ in source language

ISCED distinguishes between seven levels of education ranging from pre-primary to tertiary. Secondary education covers ages 11 or 12 through 18 or 19 and is divided into two levels: lower and upper secondary (levels 2 and 3). The English term maps quite well to ISCED

⁴ Unlike terminologies, which are only interested in real language use, multilingual thesauri may use synthetic terms or descriptive phrases to bridge lexical gaps, since they provide a reference point for other related terms that can be used to aid search and browsing, if not indexing.

levels 2 and 3, as shown in the scope note of the English term (Figure 3). However, in contrast to English, the educational systems for many target languages are structured differently with respect to these levels. For example, when mapping the Greek school system to ISCED educational levels a problem arises since the term ‘secondary’ is only used in relation to education levels in Greek, not schools (see Figure 4). Thus in ELSST, the synthetic term “ΣΧΟΛΕΙΑ ΔΕΥΤΕΡΟΒΑΘΜΙΑΣ ΕΚΠΑΙΔΕΥΣΗΣ” (literally, schools for secondary level education) is used as the Greek equivalent of SECONDARY SCHOOLS, with “ΓΥΜΝΑΣΙΑ” (corresponding to secondary schools of ISCED level 2) and “ΛΥΚΕΙΑ” (corresponding to secondary schools of ISCED level 3) as its synonyms.

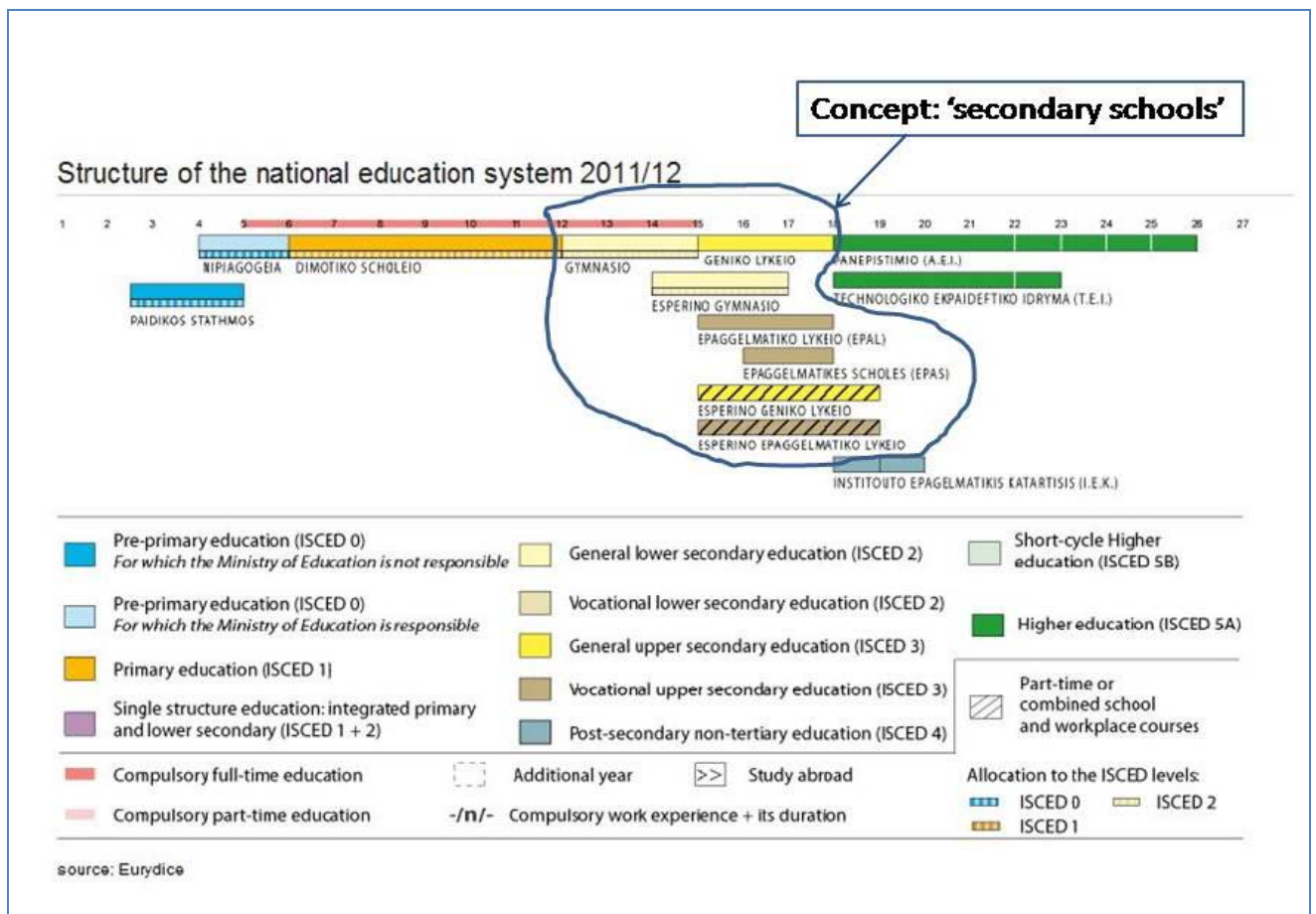


Figure 4. An example of mapping ISCED Levels 2 & 3 to the concept ‘Secondary Schools’

Besides the conceptual and linguistic difficulties in mapping ELSST with ISCED another problem lay in the scope for which ISCED was constructed: “The ISCED was not specifically developed for social survey research and facilitating comparative sociological investigations, but rather for statistical reporting, monitoring and information on national educational policy” (Schneider, 2008:316). ELSST on the other hand was designed for research and consequently uses research terminology.

Other problems with ISCED noted by researchers (e.g. Ganzeboom, 2008) include an insufficient number of categories, a very slow updating schedule, and the fact that it is based on too small a number of educational systems.

5.2 Terminologies

Multilingual term banks were consulted when looking for term definitions and/or multilingual term equivalents in ELSST. An example is EURODICAUTOM, the term bank of the European Commission which is now integrated into the Interactive Terminology for Europe (IATE) (European Union, 2012b) term bank.

5.2.1 EURODICAUTOM

EURODICAUTOM was the multilingual term bank of the European Commission's translation service, covering all EU language pairs, although not all terms were available in all languages. It covered many different subject domains encompassing the fields of activity of the European Union. Terms were often, but not necessarily, provided with a definition. While this often helped to clarify the term in the source language, it was not always enough to help find a functionally equivalent term in the target language.

An example of where EURODICAUTOM was not helpful in finding a target language equivalent is the English term INNER CITIES. In ELSST, the English term has the scope note: *“The central area of a city, especially if dilapidated or characterised by overcrowding, poverty, etc. (OED)”*

This social phenomenon does not transfer to France, where it is the suburbs, rather than the inner cities which tend to suffer social deprivation. EURODICAUTOM gave a range of French equivalents for the English term INNER CITY, including the literal CENTRE-VILLE, without capturing the connotation difference between the terms in the two languages. The solution adopted in ELSST is to leave the English term un-translated in the French version of ELSST, but to provide it with French synonyms (QUARTIERS DESHERITES, QUARTIERS PAUVRES and VIEUX QUARTIERS – literally DEPRIVED AREAS, POOR PARTS OF TOWN and OLD PARTS OF TOWN) and a translation of the English scope note.

Preferred Term
<u>INNER CITIES</u>
Scope Note
ZONE CENTRALE D'UNE VILLE, SURTOUT SI ELLE SOUFFRE DE DELABREMENT, SURPEUPLEMENT, PAUVRETE, ETC.
Synonyms/UFs
QUARTIERS DESHERITES QUARTIERS PAUVRES VIEUX QUARTIERS

Figure 5. Entry for French equivalent of INNER CITIES in ELSST

In general, EURODICAUTOM was of limited use, due to the following factors:

- gaps in coverage – many ELSST terms are phrasal, and were not found in EURODICAUTOM
- difficulty in choosing between closely related sub-domains



- lack of clarity concerning term meanings - even where terms had definitions it was not always possible to establish whether they were good equivalents
- lack of information about meaning connotations

On the plus side, the source of terms was frequently provided, which is useful, especially in the case of standardised terminology. The wide variety of near-synonyms in EURODICAUTOM provided useful suggestions for non-preferred terms in ELSST. To compensate for the lack of coverage, automatic term extraction could be investigated (Cabr Castellv , Estop Bagot, & Palatresi, 2001).

5.3 Other thesauri

Various related thesauri were consulted during the construction of ELSST, including the following: EUROVOC (European Union, 2012a), the UNESCO thesaurus (UNESCO, 2012) and the ILO thesaurus (ILO, 2012) of the European Commission, UNESCO and International Labour Organization (ILO) respectively.

The UNESCO thesaurus covers the domains of education, culture, natural sciences, social and human sciences, communication and information and is used within the organisation and beyond. EUROVOC focuses on the law and legislation of the European Union (EU) and is used in the Library of the European Parliament, the Publication Office as well as other information institutions of the EU. The ILO thesaurus focus is on topics related to economic and social development. All three thesauri are used for indexing and retrieval. ILO is reportedly also used by editors and translators as a specialised glossary.

Some more specialised thesauri were also consulted, including the thesaurus of the Education Resources Information Center (ERIC) (ERIC, 2012), used to index journal articles and other education-related materials, and the European Training Thesaurus (ETT) (Cedefop, 2012), the thesaurus of the European Centre for the Development of Vocational Training (Cedefop) which is used for indexing information relating to vocational education and training systems and programmes, policy, economical and social aspects and training-related EU actions.

While it was useful to consult other thesauri with social science coverage, differences in their scope and conceptual structure meant that it was not always possible to establish (a) whether two terms in the same language referred to the same concept in different thesauri and hence (b) whether the target language equivalent for a term in another thesaurus was suitable for the same term in ELSST. Sometimes it was only by examining the target language equivalent for a term that the difference in meaning between the source terms in two different thesauri emerged.

Term meanings in thesauri are conveyed in a number of ways, including indirectly, via their place in the hierarchy, their synonyms or related terms, or directly, via their scope notes or definitions. Disregarding other factors, Table 2 below illustrates how the terms EDUCATION, TEACHING METHODS, EDUCATIONAL SCIENCES and INSTRUCTION in six different thesauri have some overlap of meaning, based on their synonyms, especially the synonym PEDAGOGY.



ELSST	EUROVOC thesaurus	UNESCO thesaurus	European Training Thesaurus	ERIC thesaurus
EDUCATION UF: EDUCATIONAL SCIENCES, PEDAGOGY , SCHOOLING	EDUCATION	EDUCATION	EDUCATION	EDUCATION
TEACHING METHODS	TEACHING METHOD UF = PEDAGOGY	TEACHING METHODS UF: EDUCATIONAL METHODS, INSTRUCTIONAL METHODS, TEACHING STRATEGIES, TEACHING TECHNIQUES	(UF of TRAINING METHOD)	-
-	-	EDUCATIONAL SCIENCES UF: PEDAGOGY	SCIENCES OF EDUCATION UF = PEDAGOGY	-
(UF of TEACHING)	-	(UF of TEACHING)	-	INSTRUCTION UF: PEDAGOGY

Table 2. Partial equivalence of terms between thesauri based on UFs

Further investigation of these terms (by, for example, comparing their scope notes, if any, and their place in the hierarchy) was necessary to establish the degree of equivalence of the concepts expressed by these terms in each thesaurus.

An example of an ambiguity in the source term that was not apparent at a monolingual level is the English term DRUG(S) which is shown with its French equivalents in three separate thesauri in Table 3.

ELSST	ILO thesaurus	UNESCO thesaurus
DRUGS FR: DROGUES ET MEDICAMENTS	DRUG FR = DROGUE	DRUGS PHARMACEUTICALS FR = MÉDICAMENT

Table 3. Partial equivalence of terms between thesauri based on their translations

Superficially it looks as though the English term DRUG(S) is referring to the same concept in all three thesauri. However, inspection of the French language equivalents shows that this is not the case. ‘MÉDICAMENT’ (literally medicinal drugs) is the French equivalent for the term in the UNESCO thesaurus, while DROGUE (literally illegal drug) is the French equivalent in the ILO thesaurus. In ELSST, DRUGS can refer to both legal and illegal drugs, hence the decision to use DROGUES ET MEDICAMENTS (literally ILLEGAL DRUGS and MEDICINAL DRUGS) as its French equivalent.

In summary, interoperability between thesauri is limited by the following:

1. differences in scope and coverage
2. differences in conceptual framework



3. lack of term definitions
4. lack of clarity of meaning of term relationships

Problem (3) can be addressed by providing scope notes and definitions. This is particularly important in a multilingual context, and the first step in finding target language equivalents in ELSST was often to precisely define the meaning of the term in English. Problem (4) is a characteristic of many thesauri, where, for example, the BT relationship can express a number of semantic relationships, not strictly the 'isa' relationship where all objects that belong to the extension of the NT also belong to the extension of the BT. (For example, in ELSST, EDUCATIONAL PERSONNEL is the BT of TEACHERS, where it is the case that all teachers are also educational personnel. However, CHILDHOOD is the BT of CHILDREN, although children are not the same type of thing as childhood.) RT, similarly, can express a number of relationships, including cause and effect (e.g. OFFENCES in ELSST is related to DELINQUENCY and PUNISHMENT). A possible 'solution' is to convert thesauri to ontologies, where relationships between terms are more formally defined, and this is the subject of ongoing research (see for example Hahn (2003) and Biasiotti & Fernandez-Barrera(2009)).

6 Conclusion

To return to our original question about the value of re-using lexical resources in the construction of ELSST, our answer is generally affirmative, while acknowledging the difficulty of the task. Given the amount of time and effort that has gone into creating these resources in the first place, it makes sense to re-use them where possible, particularly in the case of standardised terminology. This also promotes semantic interoperability between information sources.

On the negative side, it involves a lot of work, and we encountered a variety of problems in mapping terms (via their concepts) and providing target language equivalents, due to, among other things, the differences in scope and subject coverage between lexical resources employed, the differences in conceptual framework, and the lack of definitions which are of great importance for social sciences terms.

Specifically, we conclude that the following features are desirable in any lexical resource in order to facilitate generation of target language terms, and promote its re-usability:

1. Terms must be clearly defined

Term definitions or scope notes, although not imperative in monolingual resources, are essential in multilingual resources, otherwise mapping errors can arise. Definitions also lead to a better (and more re-usable) end-product. However, although scope notes facilitate concept mapping, they are not adequate in all cases –that is, concept analysis is sometimes required. This raises doubts about the feasibility of fully automating the process of aligning lexical resources, an area of current research.

2. Terms must be unambiguous

This is a general desideratum for thesaurus construction, though not always achievable. As discussed, the process of finding a target equivalent for a term often sheds light on the nature of the underlying concept that was not apparent in the source language, and can thus help clarify the concept. The greater number of target languages, therefore, the greater, potentially, is the clarity of the concepts.



ANSI/NISO Z39.19-2005, (2005) notes that these two conditions are key to interoperability in controlled vocabularies. However, we found that even given these two conditions, exact equivalence between terms is not always possible. This is particularly evident in multilingual thesauri, where there may be cultural differences between concepts.

Acknowledgements

We would like to thank all CESSDA archive colleagues who have participated in the construction of ELSST.

7 References

- Alvheim, A. (2006): *Multilingual access to data infrastructures of the European Research Area*. Research Report, CESSDA, Norwegian Social Science Data Services: Bergen, Norway.
- ANSI/NISO Z39.19-2005 (2005): *Guidelines for the construction, format, and management of monolingual controlled vocabularies*. NISO Press: Bethesda, Maryland, U.S.A.
- Athanasίου P. (2006): The Application of Multilingualism in the European Union Context. *Legal Working Paper Series*, No.2, (March). European Central Bank. <http://www.ecb.int>.
- Balkan, L. et al. (2010): European Language Social Science Thesaurus (ELSST): Issues in designing a multilingual tool for social science researchers. *COST A31 Final Conference: Categorising Human Experience: Classification in Languages and Knowledge Systems*. Paris.
- Biasiotti, M. A. & Fernandez-Barrera, M. (2009): Enriching thesauri with ontological information: Eurovoc Thesaurus and DALOS Domain Ontology of Consumer Law. *Workshop on Legal Ontologies and Artificial Intelligence Techniques, joint with 2nd Workshop on Semantic Processing of Legal Texts. In Conjunction with ICALL 2009, 12th International Conference on Artificial Intelligence and Law*. Barcelona 8-12 June. http://idt.uab.es/images/IDT_Collections/IDT_Series/IDTSeries2_LOAIT.pdf.
- Cabrè Castellví, M. T., Estopà Bagot, R. & Palatresi, J. V. (2001): Automatic term detection: a review of current systems. In *Recent Advances in Computational Terminology* (pp. 53-87). John Benjamins Publishing Company: Philadelphia, PA.
- Campbell, K. E., & Giannangelo, K. (2007): Language barrier: getting past the classifications and terminologies roadblock. *Journal of AHIMA*, 78(2).
- Cedefop (2012): *European Training Thesaurus (ETT)* European Centre for the Development of Vocational Training. <http://libserver.cedefop.europa.eu/ett/en/>
- CESSDA PPP. (2008-2010): *WP4 Synopsis*. http://www.cessda.org/project/doc/wp04synopsis_sept08.pdf
- Chouliaraki, L. & Fairclough, N. (1999): *Discourse in late modernity – rethinking critical discourse analysis*. Edinburgh University Press: Edinburgh.
- DDC. (1989): *Dewey Decimal Classification -DDC20*. Forest Press: New York.
- DDI Alliance. (2009): *What is DDI?* <http://www.ddialliance.org/what>
- Doerr, M. (2001): Semantic problems of thesaurus mapping. *Journal of Digital Information*, 1(8).
- ERIC (2012): *ERIC thesaurus*. European Resources Information Centre http://www.eric.ed.gov/ERICWebPortal/resources/html/thesaurus/about_thesaurus.html
- European Union (2012a): *Eurovoc thesaurus*. <http://eurovoc.europa.eu/>
- European Union (2012b): *Interactive Terminology for Europe (IATE)* term bank. <http://iate.europa.eu/>



- Eurostat (1999): *Inventory of International Statistical Classifications* (1999 ed., Vol. Theme 2: Economy & Finance). European Communities: Luxembourg.
- Eurydice (2012): *European Glossary on Education* (2nd ed., 2: Educational Institutions). Eurydice, the Information Network on Education in Europe: Brussels.
- Fidrmuc, J., Ginsburgh, V. and Weber, S. (2007): Ever Closer Union or Babylonian Discord? The Official-language Problem in the European Union, *Discussion paper*, Centre for Economic Policy Research. <http://www.fidrmuc.net/research/FGW2.pdf>
- Ganzeboom, H. B. (2008): Harmonising education and occupation in cross-national comparative research. *CESSDA PPP Expert Workshop on Harmonisation Issues in Comparative Social Surveys*. Paris, France.
- Granda, P., Kramer, S. & Linnerud, J. (2008): Controlled vocabulary. In K. M. Stefan Kramer (ed.) DDI Alliance.
- Guarino, N. (1995): Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human Computer Studies*, 43 (5-6): 625-640.
- Guédon, J. C. (2008): Open access: restoring the Republic of Science and its Great Conversation. *International Conference : Open Access Infrastructures: The Future of Scientific Communication*. Athens: National Hellenic Research Foundation.
- Hahn, U. (2003): Turning informal thesauri into formal ontologies: a feasibility study on biomedical knowledge re-use, comparative and functional genomics. *Comp Funct Genom*, 4: 94-97(DOI: 10.1002/cfg.247).
- Harbsmeier, C. (2007): *Concepts that make multiple modernities: the conceptual modernisation of China in a historical and critical perspective*. <http://www.hf.uio.no/ikos/english/research/projects/tls/publications/CONCEPTS-THAT-MAKE-HISTORY%5B1%5D.pdf>
- Hayek, F. A. (1976): *The sensory order: an inquiry into the foundations of theoretical psychology*. The University of Chicago Press: Chicago, IL.
- Hodge, G. (2000): *Systems of knowledge organisation for digital libraries: beyond traditional authority files*. The Digital Library Federation Council on Library and Information Resources: Washington DC.
- ILO (2011): *International Standard Classification of Occupations (ISCO)*. International Labour Organization <http://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm>
- ILO (2012): *ILO thesaurus*. International Labour Organization <http://www.ilo.org/thesaurus>
- ISO (1985): *Documentation guidelines for the establishment and development of multilingual thesauri*. International Standard -ISO(5964). International Standardisation Organisation.
- ISO (1986): *Documentation Guidelines for the Establishment and Development of Monolingual Thesauri*. International Standard -ISO(2788). International Standardisation Organisation: Geneva.
- ISO (1990): *Terminology - vocabulary ISO (1087)*. International Standardisation Organisation: Geneva.
- ISO (2009): *ISO 704:2009 Terminology work - principles and methods*. International Standardisation Organisation: Geneva. ftp://ftp.omg.org/pub/sbvr_rtf/ISOStandards/ISO704_2009.pdf
- Kappi, C. (2009): Diachrony of scientific classifications: the case of change in social classification systems. *Address at Conference: "The Diachrony of Classification Systems"*. 12-13 March, Wassenaar, NL: COST A-31.
- Lauser, B. et al. (2008): Comparing human and automatic thesaurus mapping. *Proc. Int'l Conf. on Dublin Core and Metadata Applications 2008*, <http://edoc.hu-berlin.de/conferences/dc-2008/lauser-boris-43/PDF/lauser.pdf>.



- Luttermann K. (2011): Cultures in Dialogue. Institutional and Individual Challenges for EU Institutions and EU Citizens from the Perspective of Legal Linguistics. *Hermes – Journal of Language and Communication Studies*, No. 46.
- Miller, K. (2004): No Longer Lost in Translation. *IASSIST 2004 -Data Futures: Building on 30 Years of Advocacy*. IASSIST: Madison, Wisconsin USA.
http://www.iassistdata.org/downloads/2004/e1_miller.pdf
- Miller, K.& Matthews, B. (2001): Having the right connections: the LIMBER project. *Journal of Digital Information*, 1(8).
- Nuopponen, A. (2010a): Methods of concept analysis - a comparative study. *LSP Journal*, 1(1): 4-12.
- Nuopponen, A. (2010b): Methods of concept analysis - towards systematic concept analysis. *LSP Journal*, 1(2):5-14.
- Nuopponen, A. (2011): Methods of concept analysis - tools for systematic concept Analysis. *LSP Journal*, 2(1): 4-15.
- OECD (1999): *Classifying Educational Programmes Manual for ISCED-97 Implementation in OECD Countries* Organisation for Economic Co-operation and Development.
- Priss, U. (1996): *Relational Concept Analysis: Semantic Structures in Dictionaries and Lexical Databases*. Technischen Universität Darmstadt, Fachbereich Gesellschafts- und Geschichtswissenschaften. Darmstadt: <http://www.upriss.org.uk/papers/diss.pdf>.
- Priss, U. (2006): Formal concept analysis in information science. *Annual Review of Information Science and Technology, ASIST*, 40.
- Schneider, S. L. (Ed.) (2008): *The International Standard Classification of Education (ISCED-97)*. Mannheimer Zentrum für Europäische Sozialforschung: Mannheim, Germany.
- Schneider, S. L. (2008): Suggestions for the cross-national measurement of educational attainment: refining the ISCED-97 and improving data collection and coding procedures. In S. L. Schneider (ed.), *The International Standard Classification (ISCED-97). An evaluation of content and criterion validity for 15 European countries* (pp. 311-330). Mannheimer Zentrum für Europäische Sozialforschung (MZES): Mannheim.
- The Library of Congress. (2009, 8 10): *Library of Congress Classification*.
<http://www.loc.gov/catdir/cpsol/lcc.html>
- Tudhope, D., Koch, T., & Heery, R. (2006): *Terminology services and technology: JISC state of the art review*. Project Report: Joint Information Systems Committee, Opus: University of Bath Online Publication Store.
http://opus.bath.ac.uk/23563/1/terminology_services_and_technology_review_sep_06.pdf
- UNESCO (2006): *International Standard Classification of Education (ISCED-97)*. United Nations Educational, Scientific and Cultural Organization.
- UNESCO (2012): *UNESCO thesaurus* United Nations Educational, Scientific and Cultural Organization. <http://www2.ulcc.ac.uk/unesco/>
- University of Essex (2012): *Humanities and Social Science Electronic Thesaurus (HASSET)*. UK Data Archive, University of Essex
<http://www.data-archive.ac.uk/find/hasset-thesaurus>
- van Hage, W. R. et al. (2008): The OAEI food task: an analysis of a thesaurus alignment task in *Applied Ontology*, 1(1). IOS Press <http://www.cs.vu.nl/~guus/papers/Hage10d.pdf>.
- WHO(1998): *The World Health Report 1998 -Life in the 21st century; A Vision for All*, WHO Director General.
- WHO (2010a): *International Classification of Diseases (ICD)*. World Health Organization.
<http://www.who.int/classifications/icd/en/>



- WHO (2010b): *International Classification of Functioning Disability and Health (ICF)*. World Health Organization. <http://www.who.int/classifications/icf/en/>
- Zeng, M. L. & Hodge, G. (2011): Developing a Dublin Core application profile for the Knowledge Organisation Systems (KOS) resources. *ASIS&T Bulletin*, April/May. http://www.asis.org/Bulletin/Apr-11/AprMay11_Zeng_Hodge.html.
