Sustainable European Multilingual Vocabularies: A Model for Cooperation in Metadata Management among European Data Archives

Suzanne Barbalet

Senior Discovery Officer, English-Language Vocabularies

Sharon Bolton

Data Publishing and Curation Manager

IASSIST 2019

Sydney, Australia

Overview

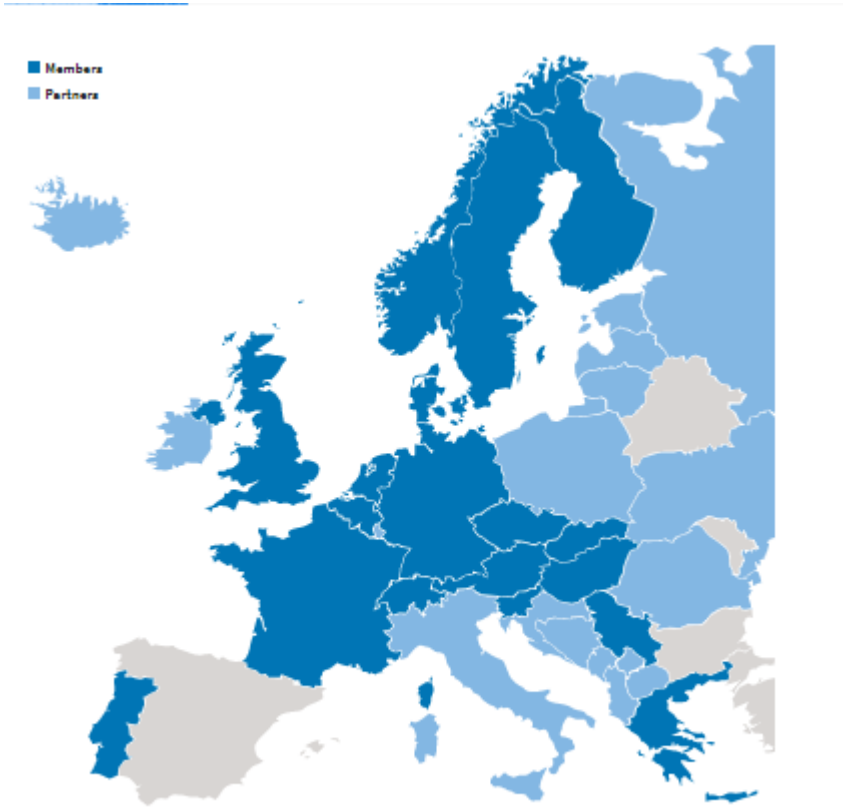
1. CESSDA and co-operative metadata management
2. The challenge of cross border vocabulary building
3. What do sustainable vocabularies look like?
4. The rewards of cross border vocabulary building
5. Lessons learnt
6. Planning for the future

UK Data Service

UK • DATA ARCHIVE

University of Essex

CESSDA-ERIC: Structure and membership



UK Data Service

UK • DATA ARCHIVE

University of Essex

CESSDA-ERIC: Mission and Vision

"The mission of CESSDA is to provide a full scale sustainable research infrastructure enabling the research community to conduct high-quality research in the social sciences contributing to the production of effective solutions to the major challenges facing society today and to facilitate teaching and learning in the social sciences."

UK Data Service

UK • DATA ARCHIVE

University of Essex

As a service provider:

The UK Data Service (UKDS) has worked together with:

Finnish Social Science Data Archive (FSD)

Norwegian Centre for Research Data (NSD)

Leibniz Institute for the Social Sciences (Gesis)

Swiss Centre of Expertise in the Social Sciences (FORS)

UK Data Service

UK · DATA ARCHIVE

University of Essex

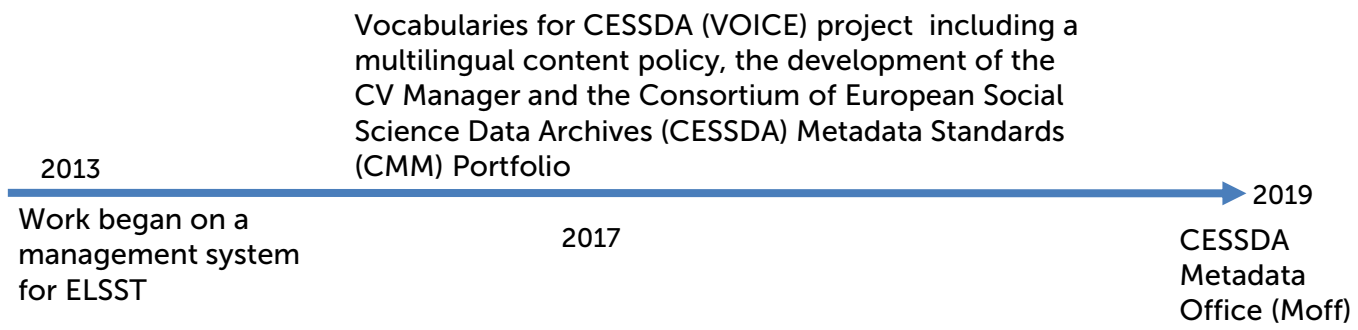
All CESSDA members have contributed to:

- the development of English Language Social Science Thesaurus (ELSST)
- the development of the CV Manager
- the development of the CESSDA Metadata Standards (CMM) Portfolio
- assisted with testing through various phases of development of the Thesaurus Management System

UK Data Service UK • DATA ARCHIVE University of Essex

CESSDA-ERIC Vocabularies: – A Model of Cooperation between European Social Science Data Archives

Timeline of vocabularies projects 2013 - 2019



UK Data Service

UK • DATA ARCHIVE

University of Essex

CESSDA metadata products

A sustainable research infrastructure

requires

sustainable metadata

UK Data Service

UK • DATA ARCHIVE

University of Essex

The challenge: building vocabularies across languages and social science disciplines

- concept formation
- provenance of concept
- early identification of translation issues
- access to similar data

Concept formation

Identifying the concept and translating the concept is most often a positive and enriching interactive process.

BUT

when one word derives it meaning in context in the source language but not in other languages

AND

that concept is a significant social problem such as DRUG USE

We had what we referred to as the DRUGS problem (in more ways than one) as a number of languages had separate terms for medicinal and recreational drugs.

Concepts and variables

Academic concepts tend to be *international* in scope

AND

social science and humanities disciplines are no exception

BUT

variables chosen to measure concepts can be culture dependent

Provenance of concepts

- a large pool of data references helps negotiate provenance issues that cross disciplinary concepts can pose
- cooperative metadata management draws on a pool of expertise and data resources to establish provenance
- provenance research is time consuming and best shared

Identification of translation issues

- while translation issues are addressed in meetings there is often a time lapse before translation begins
- those translating are not necessarily those attending vocabulary construction meetings and may not always be familiar with the data the concept is designed to retrieve
- legacy translation issues are harder to resolve in thesauri once the concept is situated in structures and semantic relationships established
- these issues are addressed through regular translator training sessions conducted by Finnish Social Science Data Archive (FSD) and Lorna Balkan, UK Data Service

UK Data Service UK • DATA ARCHIVE University of Essex

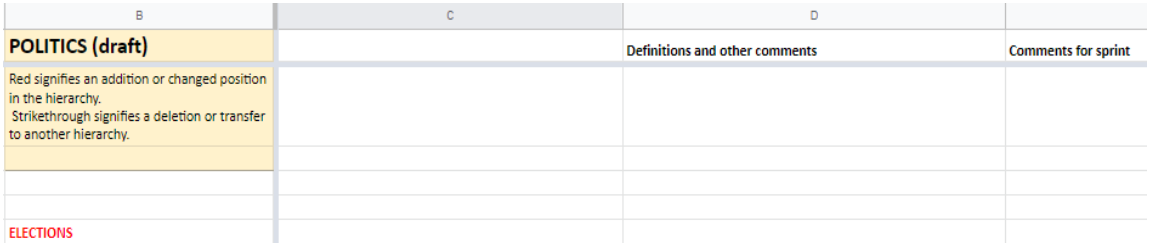
Cross border retrieval

- data examples are shared in sprint and translators' meetings
- additions to the vocabularies go through many stages of scrutiny
- translation issues are reviewed

Some examples

Current work on ELSST: POLITICS hierarchy

The draft was prepared by FSD staff who have expertise in this subject area



B	C	D	
POLITICS (draft)		Definitions and other comments	Comments for sprint
Red signifies an addition or changed position in the hierarchy. Strikethrough signifies a deletion or transfer to another hierarchy.			
ELECTIONS			

Preparation work done by each member of the team prior to meeting included sourcing scope notes and providing data examples

UK Data Service

UK · DATA ARCHIVE

University of Essex

Further examples: ELSST

- earlier work on the restructuring of CRIME and LAW terms was guided by UKDS expertise (Dr Sharon Bolton – Criminologist)

- work on creating a WELL-BEING hierarchy was guided by an expert panel comprising subject experts from both staff designing and managing the collection of this data in the *Understanding Society* and the *British Household Panel Survey,* as well as a UKDS staff member with relevant expertise.

- one of the first ELSST revision projects we undertook together was in response to the aforementioned DRUGS terms problem

Why cooperation in vocabulary construction?

It is a world where *open data* has become an expectation

BUT

an environment too where *security* of data is paramount

WHERE

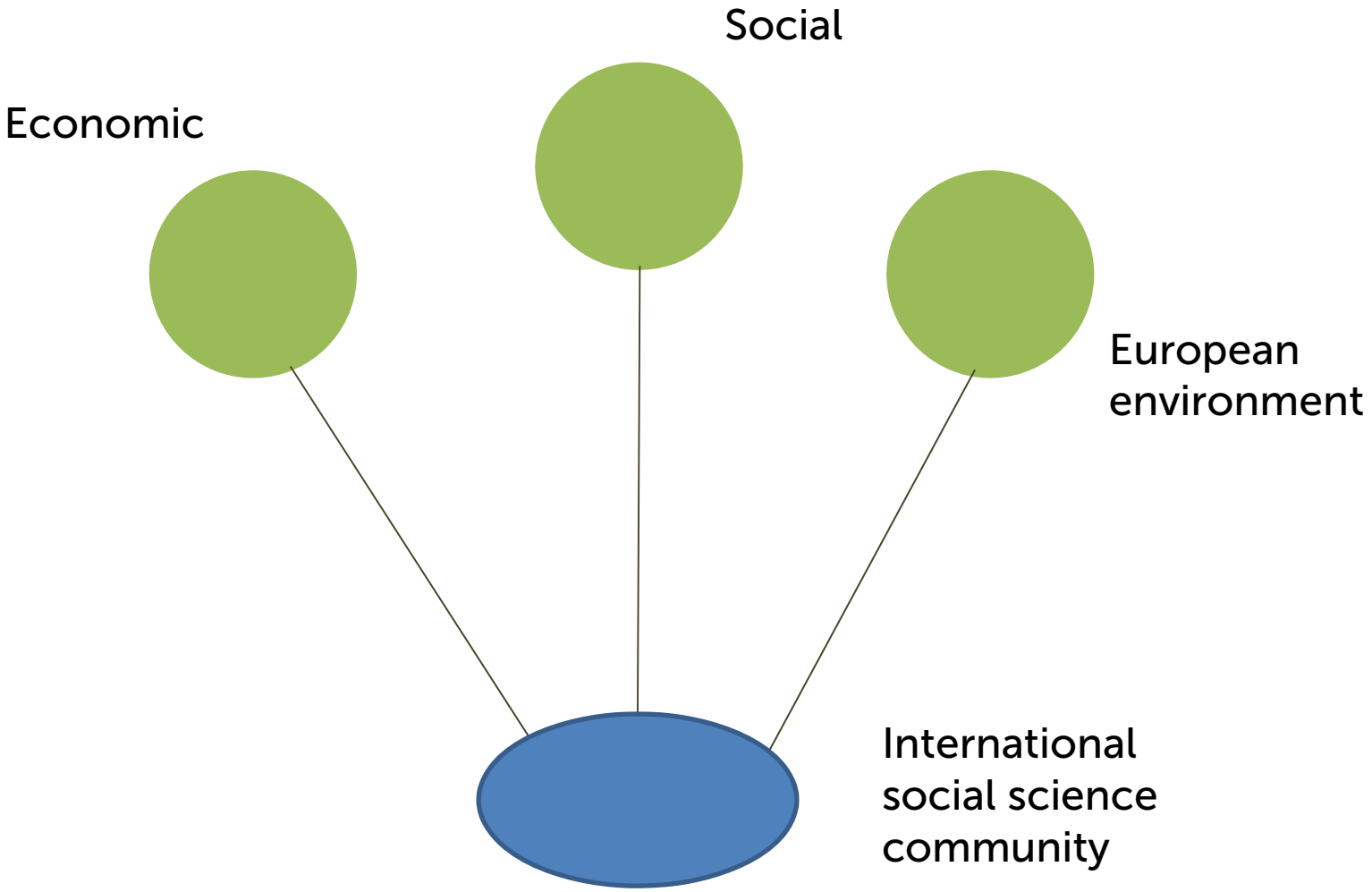
data archives share users with a similar profile, and users who carry credentials that enable them to become approved users, then collaborating in the production of metadata has many benefits.

What do sustainable vocabularies look like?

Vocabularies that:

- will ensure continuity in development
- manage change
- satisfy the needs of all contributors
- are economical in the use of resources
- plan for the future

Sustainable vocabularies



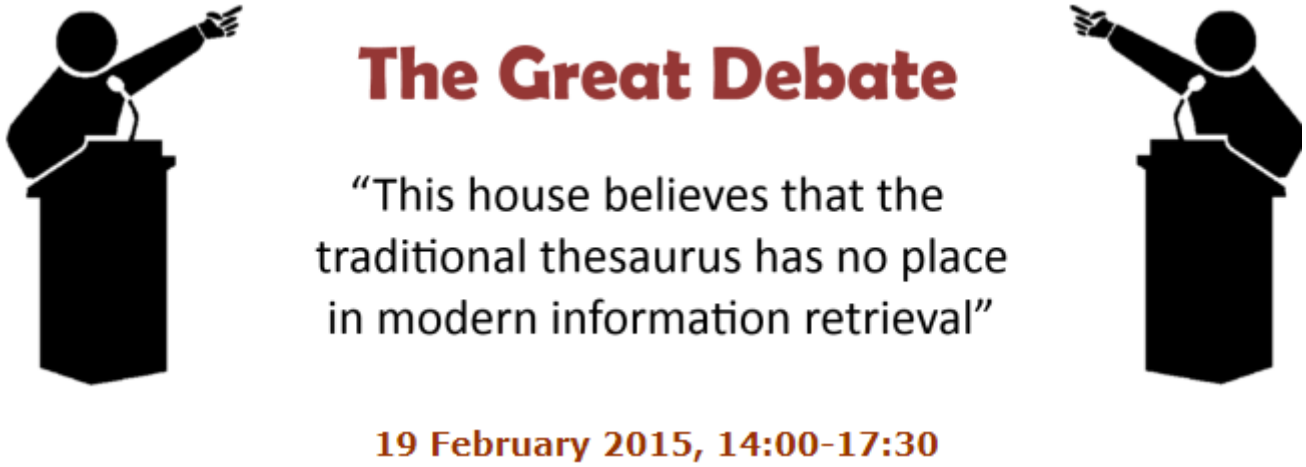
UK Data Service

UK • DATA ARCHIVE

University of Essex

Dimensions of sustainability: Economic

Consensus of experts that vocabularies are expensive to maintain



Reference: http://event-archive.iskouk.org/content/great-debate

Conclusion of this debate was that the traditional thesauri have a place in the future of information retrieval and that the UKDS thesaurus HASSET, which shares much of its vocabulary with ELSST, demonstrated the innovative qualities required.

What does cooperation in metadata management offer?

CESSDA vocabulary projects have ensured:

- high quality social science metadata via access to a pool of cross disciplinary expertise.
- revision of vocabularies necessary to keep social science metadata current
- standards maintained through the production of ‘user guide’ documentation
- access to a broad spectrum of funding resources

Lessons learnt

Unity of purpose: matching the vocabulary to the task

- a survey conducted midway through the CESSDA-ELSST project found differences in the use of ELSST. Some CESSDA partners did not use it directly for indexing.

- a SERISS project (see IASSIST 2019 Poster Session) also found differences between study indexing and question indexing indicating the importance of matching the vocabulary to the task

Sustainable or ‘slow’ vocabularies

- ‘slow’ or sustainable vocabularies does not mean inefficiency or lack of innovation
- it means revisiting problem areas for translation
- ensuring once agreement is reached that we are working within the guidelines of ISO standards
- building structures that incorporate provision for currency updates without requiring total restructuring
- building strong vocabularies that will withstand the rigor of automatic search and retrieval systems

Questions

Suzanne Barbalet : sbarba@essex.ac.uk

Sharon Bolton: sharonb@essex.ac.uk