O uso dos **vocabulários controlados** na gestão de dados em ciências sociais: o trabalho do APIS no âmbito do projeto CESSDA Metadata Office (MDO)

Patrícia Miranda
Pedro Moura Ferreira

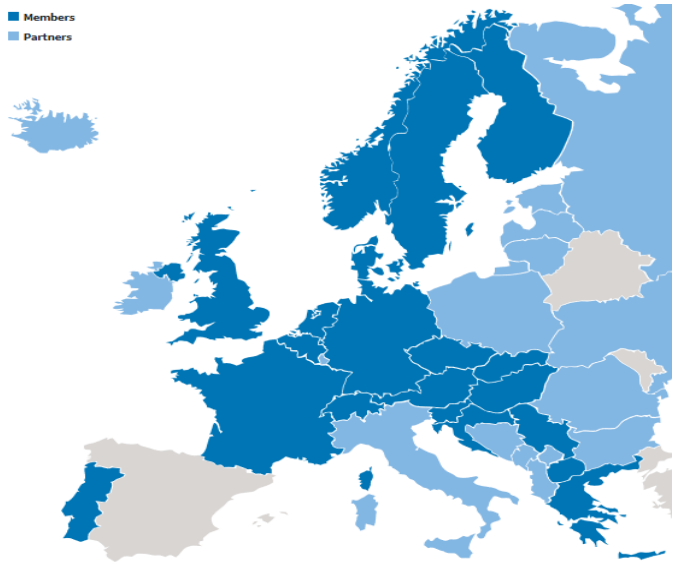
Quem somos e o que fazemos
APIS: enquadramento no PASSDA e ligação ao CESSDA

- O Arquivo Português de Informação Social (APIS) é uma das componentes do PASSDA (Production and Archive of Social Science Data).

- O PASSDA é uma infraestrutura de investigação que faz parte do Roteiro Nacional de Infraestruturas de Investigação de Interesse Estratégico apoiado pela FCT, com base numa parceria que envolve unidades e centros de investigação da Universidade de Lisboa, do ISCTE-IUL, da Universidade de Coimbra e da Universidade do Porto.

- Tem por objetivo a recolha, arquivo e disseminação de dados sobre atitudes, valores e comportamentos sociopolíticos.

- A principal atividade do APIS consiste em assegurar a continuidade do acesso aos dados produzidos pelos projetos de investigação, através de um conjunto de atividades de curadoria, preservação e disseminação.



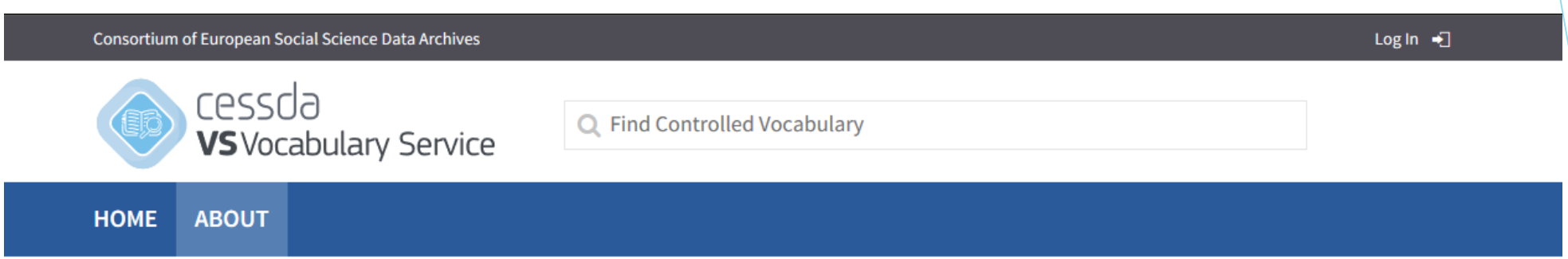
- O APIS é membro da infraestrutura europeia CESSDA ERIC (Consortium of European Social Science Data Archives - European Research Infrastructure Consortium).
- O CESSDA fornece serviços de dados em larga escala para as ciências sociais.

  Reúne os arquivos de dados das ciências sociais de toda a Europa, com o objetivo de apoiar a investigação e a cooperação nacionais internacionais.

CESSDA ⟷ Working Group **MDO** (Metadata Office)

- O APIS participa no CESSDA WG MDO, que tem por objetivo o desenvolvimento de metadados standardizados e a criação do CESSDA Metadata Model (CMM).
- O CMM baseia-se no Data Documentation Initiative (DDI).

"The objective of the CMM project is to develop a standardised metadata design and practice for CESSDA. The outcome will be a Metadata Standards Portfolio that will support resource discovery and question banks. The Portfolio will be compliant with the ***Data Documentation Initiative (DDI)*** *international standard for describing statistical and social science data and will include a* ***core metadata model*** *and* ***controlled vocabularies*** *for relevant metadata fields."* (from CESSDA website)



CESSDA Vocabulary Service enables users to discover, browse, and download controlled vocabularies in a variety of languages.
The service is provided by the Consortium of European Social Science Data Archives (CESSDA).

The majority of the source (English) vocabularies included in the service have been created by the DDI Alliance. The Data Documentation Initiative (DDI) is an international standard for describing data produced by surveys and other observational methods in the social, behavioural, economic, and health sciences.
The language versions of the DDI vocabularies have been provided by CESSDA member organisations in different countries, as they use the vocabularies to describe research data.

The service also contains an Editor, where authorised users create, manage and translate the vocabularies. Access to the Editor is restricted.

CV – Controlled Vocabulary

- Vocabulário controlado (controlled vocabulary) foi descrito como: "*A set of subject terms, and rules for their use in assigning terms to materials for indexing and retrieval.*". https://repository.arizona.edu/bitstream/handle/10150/105435/glossary.html

- Nos vocabulários controlados, um termo consiste numa ou mais palavras usadas para representar um conceito. Os termos são selecionados/construídos a partir do idioma "natural"/de origem para inclusão num "vocabulário controlado" estabelecido.

CVs ↔ Metadados

- Os **vocabulários controlados** (CVs) são listas de termos usados para os **metadados** e outras informações.

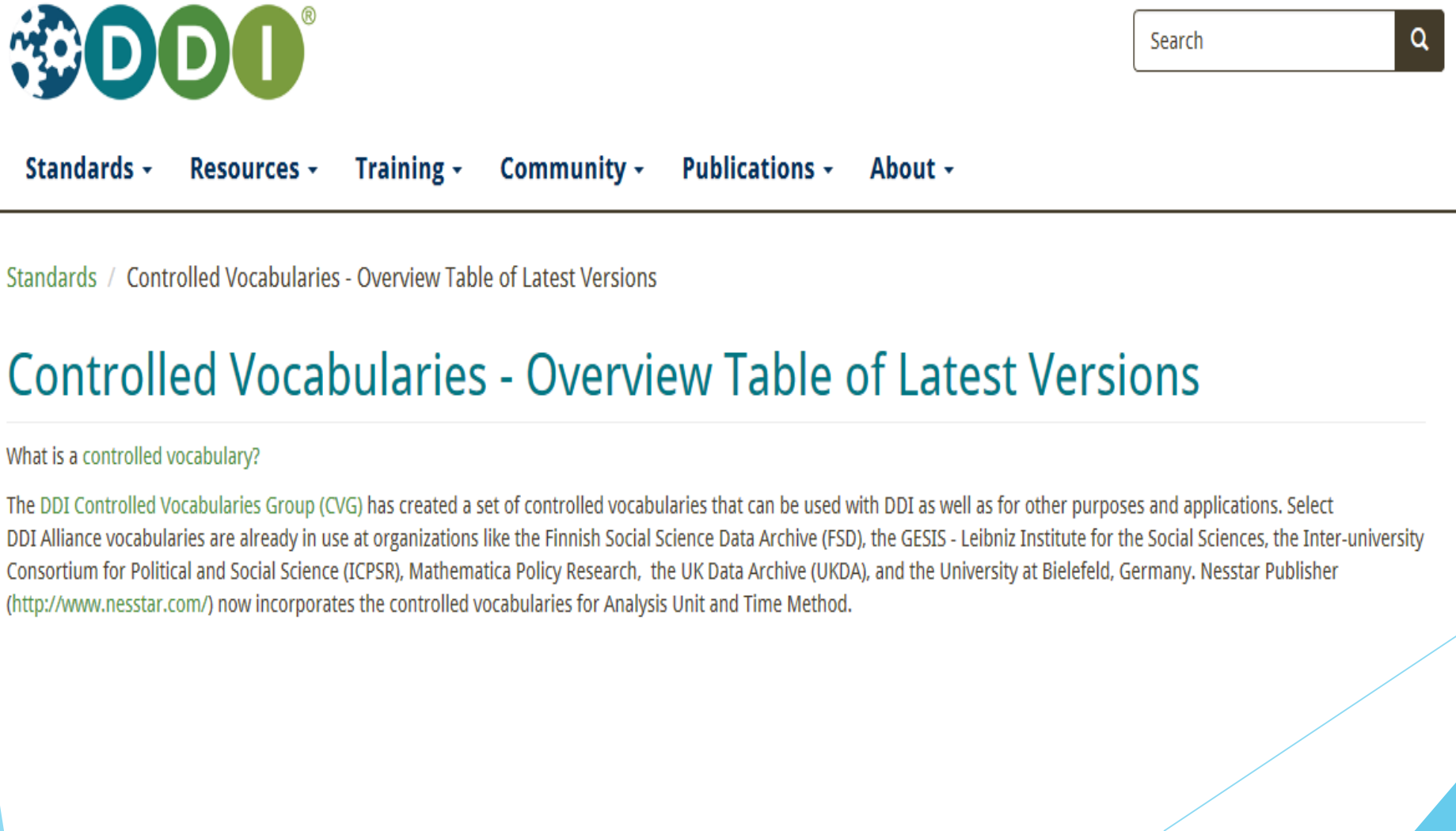
- Podem ser listas curtas e simples, ou construções mais complexas, incluindo relações de equivalência, relações hierárquicas, etc.
- Exemplos: LCSH (Library of Congress Subject Headings); thesaurus multilingue do **ELSST** (European Language Social Science Thesaurus) usado por arquivos de dados europeus.

CVs ↔ DDI

- O **DDI (Data Documentation Initiative)** é um padrão internacional para descrever os dados produzidos nas ciências sociais, comportamentais, económicas e de saúde. O DDI é um padrão gratuito que pode documentar e gerenciar diferentes estágios no ciclo de vida dos dados da pesquisa, como conceptualização, recolha, processamento, disseminação, descoberta e arquivo.

- A documentação de dados com o DDI facilita a sua compreensão, interpretação e utilização - por pessoas, sistemas de software e redes de computadores.

- https://ddialliance.org/controlled-vocabularies

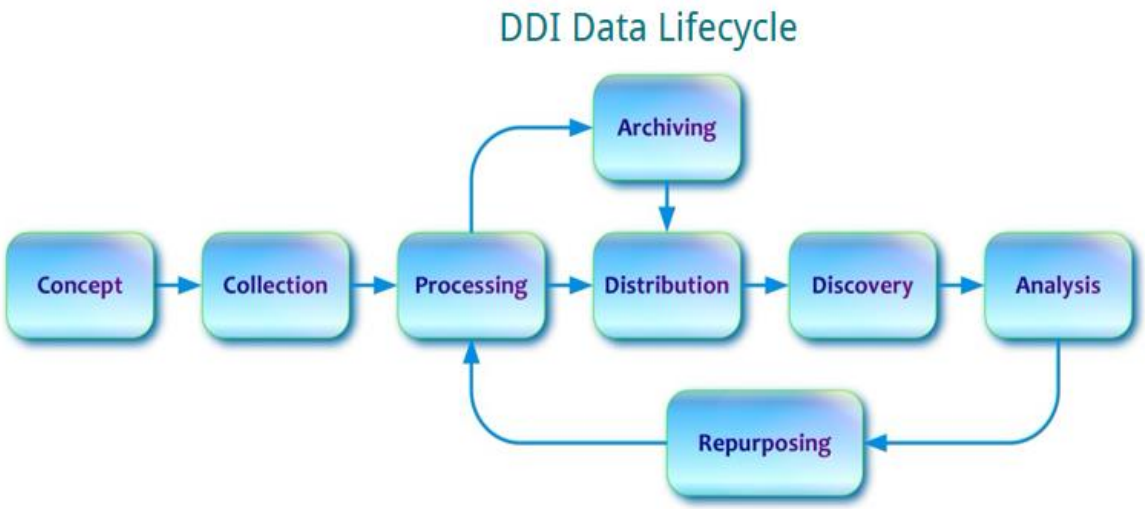


Why Use DDI?

DDI encourages **comprehensive description** of data for discovery and analysis and supports **effective data sharing**. Because DDI is a **structured** standard, it facilitates **machine-actionability and interoperability** and it can actually be used to **drive systems**. Another feature of DDI is its focus on **metadata reuse**; "enter once, use often" means you can reuse metadata over the course of the data life cycle to avoid costly duplication of effort.

DDI has advantages for several different audiences:

+ Librarians
+ Managers
+ Repositories
+ Researchers
+ Developers



Vocabulários controlados: onde e como são criados e/ou usados

- Idealmente, os termos incluídos num vocabulário controlado devem ser exaustivos (abrangendo toda a dimensão da questão), mutuamente exclusivos (sem sobreposições entre termos) e claramente definidos (definições e notas fornecidas para os significados dos termos).

- Os CVs são frequentemente usados em contextos específicos e as definições/notas esclarecem e desambiguam o significado de um termo num contexto específico, pois pode diferir do significado na linguagem "natural" (*source language*).

Os vocabulários controlados no contexto do CESSDA – Metadata Office (MDO)

- O CESSDA CV manager é uma ferramenta que permite a criação, versão e manutenção de vocabulários controlados, sua tradução para os idiomas dos países membros e acesso a todos os CVs.
- Este projeto pretende produzir uma ferramenta para o CESSDA gerenciar vocabulários controlados (CVs), pois eles constituem parte crucial do padrão de metadados do CESSDA.
- O APIS está envolvido neste projeto, sendo responsável pela tradução dos vocabulários controlados do CESSDA/DDI para a língua portuguesa.

- Link para os vocabulários controlados (CESSDA/DDI): https://vocabularies.cessda.eu/#!discover

- Os vocabulários são usados para ajudar na organização e recuperação de informações no CDC (CESSDA Data Catalogue) e também no Euro Question Bank.

- É importante haver um padrão/referência para os metadados no âmbito das ciências sociais, sendo o Data Documentation Initiative (DDI) usado como *standard internacional*.

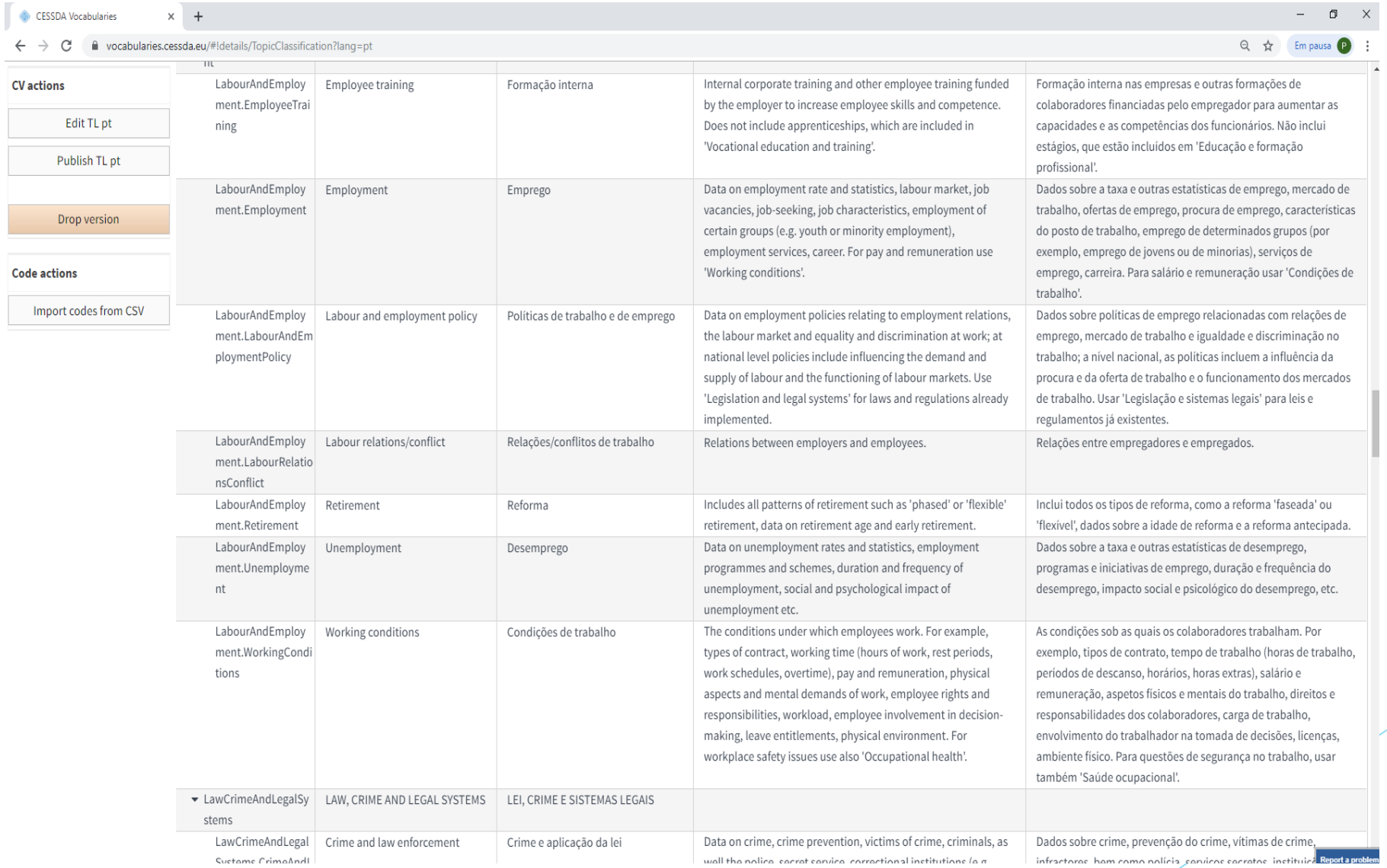
- No site do DDI Alliance há uma tabela com uma lista das últimas versões de CVs (disponíveis também no site do CESSDA), permitindo o download em vários formatos.

- …e a lista de organizações e projetos que adotaram o DDI standard continua a aumentar: https://ddialliance.org/ddi-adopters

➢ Vocabulários controlados que correspondem a metadados e que o APIS traduziu no âmbito do projeto MDO:

- Analysis Unit (DDI Alliance)
- Time Method (DDI Alliance)
- General Data Format (DDI Alliance)
- Mode Of Collection (DDI Alliance)
- Sampling Procedure (DDI Alliance)
- Topic Classification (CESSDA)
- Data Source Type (DDI Alliance)
- Type Of Instrument (DDI Alliance)
- Summary Statistic Type (DDI Alliance)
- Commonality Type (DDI Alliance)

Exemplo de tradução de CV: Topic Classification (by CESSDA)



LabourAndEmployment.EmployeeTraining	Employee training	Formação interna	Internal corporate training and other employee training funded by the employer to increase employee skills and competence. Does not include apprenticeships, which are included in 'Vocational education and training'.	Formação interna nas empresas e outras formações de colaboradores financiadas pelo empregador para aumentar as capacidades e as competências dos funcionários. Não inclui estágios, que estão incluídos em 'Educação e formação profissional'.
LabourAndEmployment.Employment	Employment	Emprego	Data on employment rate and statistics, labour market, job vacancies, job-seeking, job characteristics, employment of certain groups (e.g. youth or minority employment), employment services, career. For pay and remuneration use 'Working conditions'.	Dados sobre a taxa e outras estatísticas de emprego, mercado de trabalho, ofertas de emprego, procura de emprego, características do posto de trabalho, emprego de determinados grupos (por exemplo, emprego de jovens ou de minorias), serviços de emprego, carreira. Para salário e remuneração usar 'Condições de trabalho'.
LabourAndEmployment.LabourAndEmploymentPolicy	Labour and employment policy	Políticas de trabalho e de emprego	Data on employment policies relating to employment relations, the labour market and equality and discrimination at work; at national level policies include influencing the demand and supply of labour and the functioning of labour markets. Use 'Legislation and legal systems' for laws and regulations already implemented.	Dados sobre políticas de emprego relacionadas com relações de emprego, mercado de trabalho e igualdade e discriminação no trabalho; a nível nacional, as políticas incluem a influência da procura e da oferta de trabalho e o funcionamento dos mercados de trabalho. Usar 'Legislação e sistemas legais' para leis e regulamentos já existentes.
LabourAndEmployment.LabourRelationsConflict	Labour relations/conflict	Relações/conflitos de trabalho	Relations between employers and employees.	Relações entre empregadores e empregados.
LabourAndEmployment.Retirement	Retirement	Reforma	Includes all patterns of retirement such as 'phased' or 'flexible' retirement, data on retirement age and early retirement.	Inclui todos os tipos de reforma, como a reforma 'faseada' ou 'flexível', dados sobre a idade de reforma e a reforma antecipada.
LabourAndEmployment.Unemployment	Unemployment	Desemprego	Data on unemployment rates and statistics, employment programmes and schemes, duration and frequency of unemployment, social and psychological impact of unemployment etc.	Dados sobre a taxa e outras estatísticas de desemprego, programas e iniciativas de emprego, duração e frequência do desemprego, impacto social e psicológico do desemprego, etc.
LabourAndEmployment.WorkingConditions	Working conditions	Condições de trabalho	The conditions under which employees work. For example, types of contract, working time (hours of work, rest periods, work schedules, overtime), pay and remuneration, physical aspects and mental demands of work, employee rights and responsibilities, workload, employee involvement in decision-making, leave entitlements, physical environment. For workplace safety issues use also 'Occupational health'.	As condições sob as quais os colaboradores trabalham. Por exemplo, tipos de contrato, tempo de trabalho (horas de trabalho, períodos de descanso, horários, horas extras), salário e remuneração, aspetos físicos e mentais do trabalho, direitos e responsabilidades dos colaboradores, carga de trabalho, envolvimento do trabalhador na tomada de decisões, licenças, ambiente físico. Para questões de segurança no trabalho, usar também 'Saúde ocupacional'.
▾ LawCrimeAndLegalSystems	LAW, CRIME AND LEGAL SYSTEMS	LEI, CRIME E SISTEMAS LEGAIS		
LawCrimeAndLegalSystems.CrimeAndl	Crime and law enforcement	Crime e aplicação da lei	Data on crime, crime prevention, victims of crime, criminals, as well the police, secret service, correctional institutions (e.g.	Dados sobre crime, prevenção do crime, vítimas de crime, infractores, bem como polícia, serviços secretos, instituiç

Exemplo de vocabulário controlado associado ao metadado “unidade de análise”

Code Value	Code descriptive term
Individual	Indivíduo
OrganizationOrInstitution	Organização/Instituição
Family	Família
Family.HouseholdFamily	Família: Agregado familiar
Household	Agregado doméstico
HousingUnit	Unidade residencial
EventOrProcessOrActivity	Evento/Processo/Atividade
GeographicUnit	Unidade geográfica
PoliticalAdministrativeArea	Área político-administrativa
TimeUnit	Unidade de tempo
MediaUnit	Unidade de Media
MediaUnit.Sound	Unidade de Media: Som
MediaUnit.StillImage	Unidade de Media: Imagem
MediaUnit.Text	Unidade de Media: Texto
MediaUnit.Video	Unidade de Media: Vídeo
Group	Grupo
Object	Objeto
Other	Outro

Jaaskelainen, Taina; Moschner, Meinhard; Wackerow, Joachim. 2009. **Controlled Vocabularies for DDI 3: Enhancing Machine-Actionability**, IASSIST Quarterly, Vol 33, No 1-2

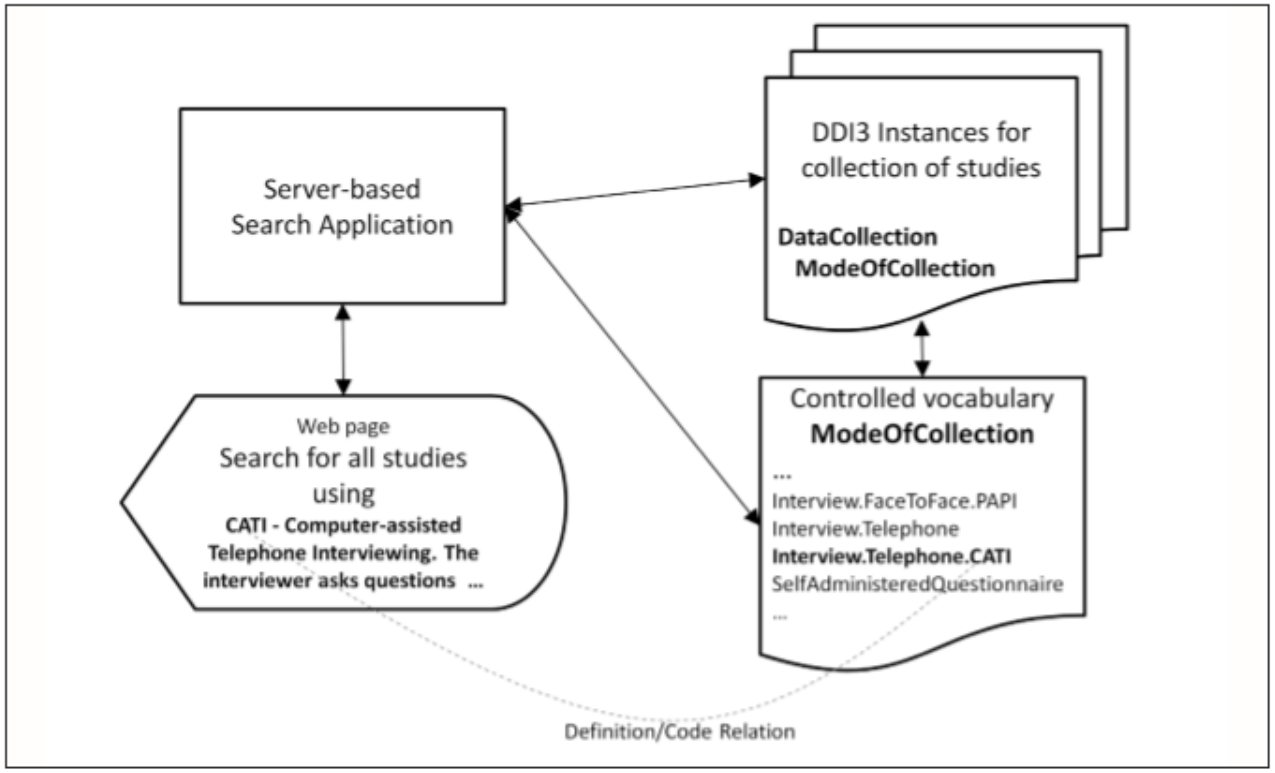
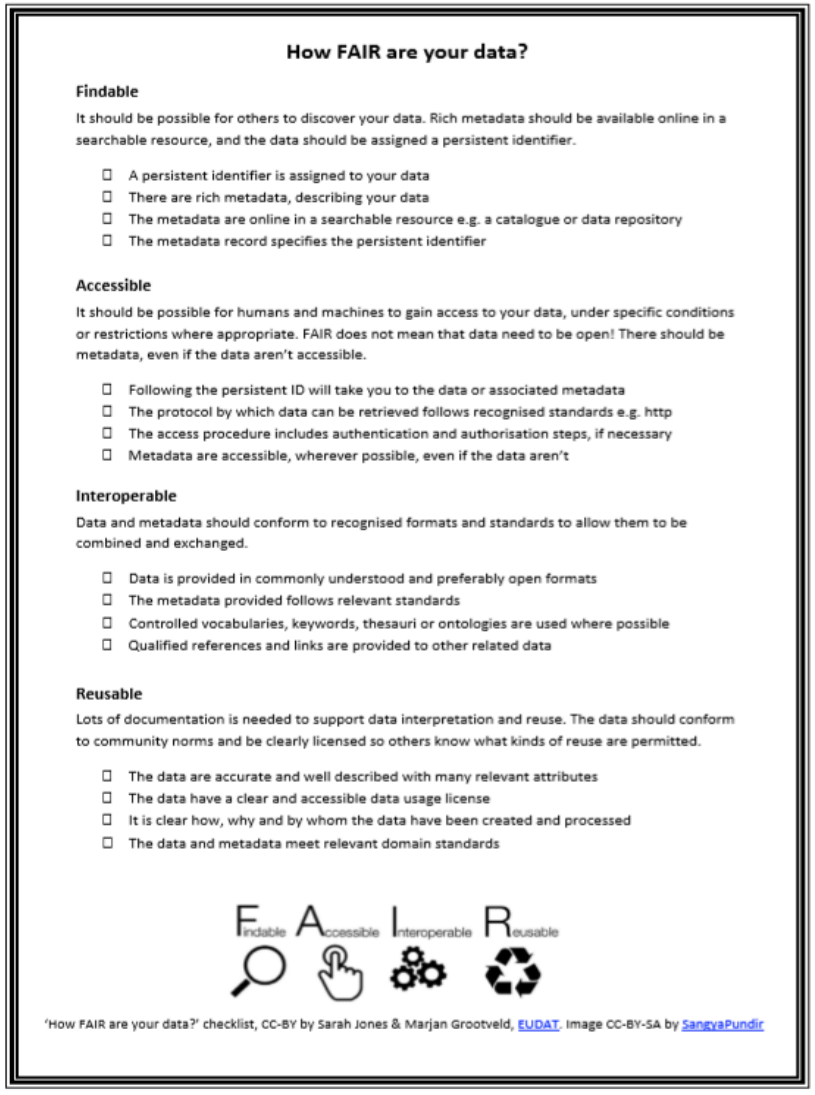


Figure 1. Filtered search enabled by controlled vocabulary

Vocabulários controlados - **FAIR**



How FAIR are your data?

Findable

It should be possible for others to discover your data. Rich metadata should be available online in a searchable resource, and the data should be assigned a persistent identifier.

- ☐ A persistent identifier is assigned to your data
- ☐ There are rich metadata, describing your data
- ☐ The metadata are online in a searchable resource e.g. a catalogue or data repository
- ☐ The metadata record specifies the persistent identifier

Accessible

It should be possible for humans and machines to gain access to your data, under specific conditions or restrictions where appropriate. FAIR does not mean that data need to be open! There should be metadata, even if the data aren't accessible.

- ☐ Following the persistent ID will take you to the data or associated metadata
- ☐ The protocol by which data can be retrieved follows recognised standards e.g. http
- ☐ The access procedure includes authentication and authorisation steps, if necessary
- ☐ Metadata are accessible, wherever possible, even if the data aren't

Interoperable

Data and metadata should conform to recognised formats and standards to allow them to be combined and exchanged.

- ☐ Data is provided in commonly understood and preferably open formats
- ☐ The metadata provided follows relevant standards
- ☐ Controlled vocabularies, keywords, thesauri or ontologies are used where possible
- ☐ Qualified references and links are provided to other related data

Reusable

Lots of documentation is needed to support data interpretation and reuse. The data should conform to community norms and be clearly licensed so others know what kinds of reuse are permitted.

- ☐ The data are accurate and well described with many relevant attributes
- ☐ The data have a clear and accessible data usage license
- ☐ It is clear how, why and by whom the data have been created and processed
- ☐ The data and metadata meet relevant domain standards

'How FAIR are your data?' checklist, CC-BY by Sarah Jones & Marjan Grootveld, EUDAT. Image CC-BY-SA by SangyaPundir

O uso de vocabulários controlados vai de encontro aos princípios **FAIR**, nomeadamente, a *interoperabilidade*, ao permitir colocar dados e metadados em formatos (re)conhecidos e padronizados.

Os vocabulários controlados, em muitos casos, aplicam-se a um domínio específico.

Aqui centrámo-nos nas ciências sociais e nos CV's do DDI Alliance e do CESSDA.

VANTAGENS dos vocabulários controlados:

- Interoperabilidade;
- Multilinguismo;
- Maior facilidade de pesquisa/recuperação de informação específica.

Implicações dos vocabulários controlados:

- Obrigatoriedade do uso de termos específicos;
- Acompanhamento das atualizações dos vocabulários controlados;
- Sincronização entre versões originais e traduzidas dos vocabulários controlados.

Obrigada pela atenção!