European Commission

Digital Skills and Knowledge Concepts:
Labelling the ESCO classification

Technical Report – October 2022

Employment, Social Affairs and Inclusion

Table of Contents

Table of Figures

Table of Charts

1. Introduction

ESCO identifies and categorises skills, competences and occupations relevant for the European labour market. It systematically describes the skills needed to perform an occupation, and it supports the provision of multilingual labour market services and the design of training programs and curricula, with the goal of bridging the communication gap between the labour market and the domain of education and training. In its current version[1], ESCO includes 13890 skills and knowledge concepts available in 28 languages.

Such a rich dataset can be further broken down by looking at the reusability level[2] of a skill or knowledge concepts or at their nature. In particular, a green label was introduced in 2022 to identify the skills needed in a green economy.

ESCO version 1.1.1 introduced the label of digital skills, thus providing an additional element in support of the twin transition. The following paragraphs describe methodology designed and adopted to label digital skills, the results of this exercise, and guides implementers in their use of ESCO digital concepts.

2. Methodology

The labelling of digital skills and knowledge concepts follows a 5-steps methodology, which combines human labelling and validation with the use of Machine Learning (ML) algorithms. Figure 1 below summarises the steps in a diagram.

[1] ESCO version 1.1.1 was published in September 2022.,

[2] ESCO distinguishes between occupation specifics, sector specifics, cross-sectoral and transversal skills and knowledge concepts

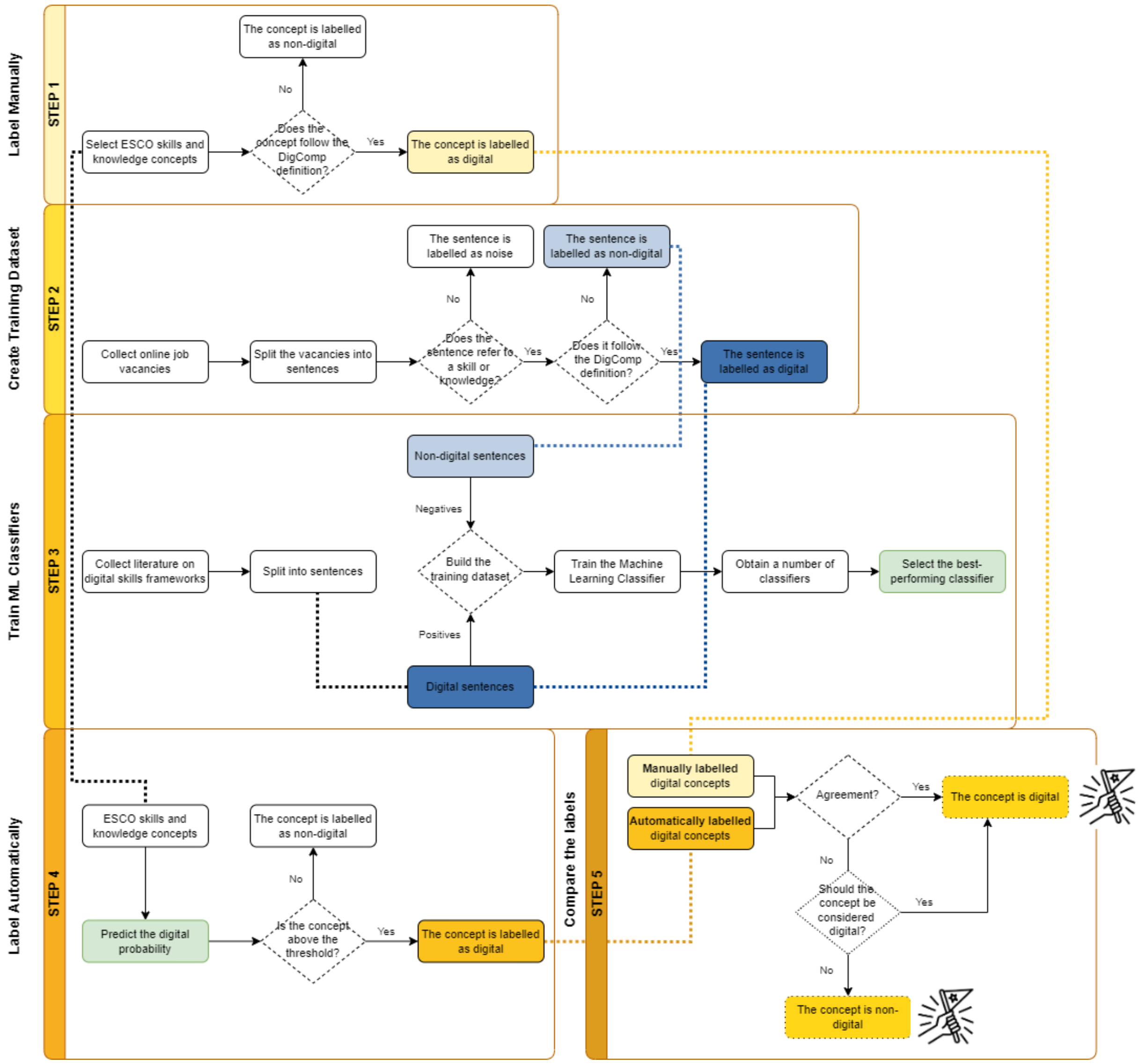


Figure 1: The five-steps methodology to label digital ESCO skills and knowledge concepts

In the **first step**, ESCO skills and knowledge concepts are manually labelled based on *The Digital Competence Framework for Citizens* (DigComp 2.2), which is a framework developed to provide a common language to identify and describe the key areas of digital competence. DigComp uses the definition of digital competence adopted in the Council Recommendation on Key Competences for Life-long Learning, which is described as follows:

Digital competence involves the confident, critical and responsible use of, and engagement with, digital technologies for learning, at work, and for participation in society. It includes information and data literacy, communication and collaboration, media literacy, digital content creation (including programming), safety (including digital well-being and competences related to cybersecurity), intellectual property related questions, problem solving and critical thinking. (Council Recommendation on Key Competences for Lifelong Learning, 22 May 2018, ST 9009 2018 INIT).

When running the manual labelling activity, each concept is analysed based on its preferred term, non-preferred terms, and description. The labelling consists in verifying whether an ESCO concept should or should not be considered as digital, following the definition above. The labelling rules are further specified in Annex 6.2.

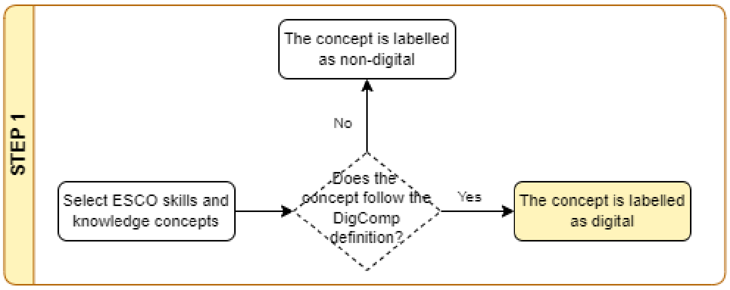


Figure 2: The first step to label digital ESCO skills and knowledge concepts

The **second step** is devoted to the manual labelling of online job vacancies (OJAs). For this task, OJAs are collected from the database of EURES, the European cooperation network of employment services. Using a multilingual text splitter developed by the ESCO team, OJAs are then divided into sentences, where each sentence has a unique identifier. Sentences are manually validated, where the validator can choose whether the sentence referred to the use of digital skills (following the DigComp definition), non-digital skills, or whether the sentence is to be considered as *noise*, meaning that it does not include information on the skills needed for the job.

The set of unique sentences was divided into two batches. For the first batch the digital labeling of sentences was done in an entirely manual manner without additional support. For the second batch suggestions in the form of probabilities were computed for each sentence, where a higher probability represents a higher likelihood of a sentence to be labelled as digital. In addition, sentences were grouped in small clusters (size of clusters ranged between 1 and 20 sentences) through a finetuned Bidirectional Encoder Representations from Transformers (BERT) language model, with the aim of speeding manual validation. Grouping sentences in clusters eased the workload to distinguish *noise* from *non-noise* sentences, where clusters composed exclusively of *noise* sentences were removed from the second batch of validation.

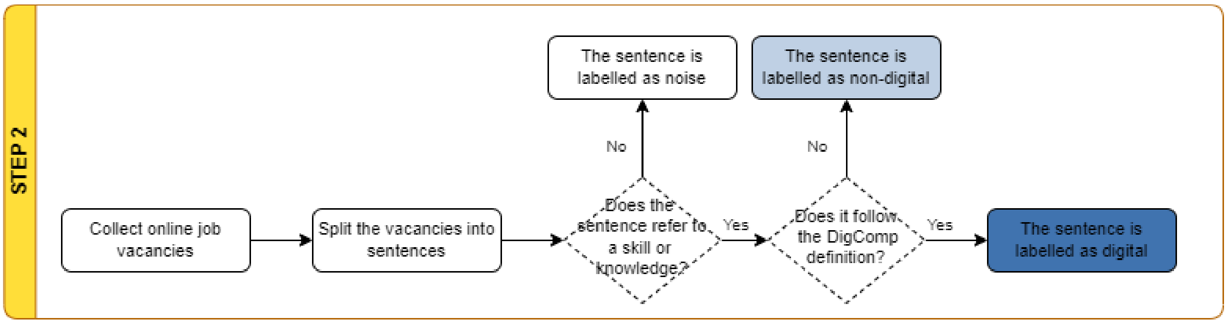


Figure 3: The second step to label digital ESCO skills and knowledge concepts

The **third step** involves the development of a machine learning (ML) model to classify ESCO skills and knowledge concepts as digital or not. This means that, given one sentence, the model is trained to assign a probability score to distinguish digital and non-digital sentences. The classifier is based on the BERT model and is fine-tuned using two sources of labour-specific information: the labelled sentences resulting from the second step described above, and a set of descriptions and examples from the literature. The literature employed for the training dataset is based on existing frameworks that describe digital skills, which are listed later in this section.

To develop the model, the data was divided in training and validation data. Different classifiers resulting from the training activity were compared, and the best-performing classifier was selected. More details on this activity are available in Annex 6.1.

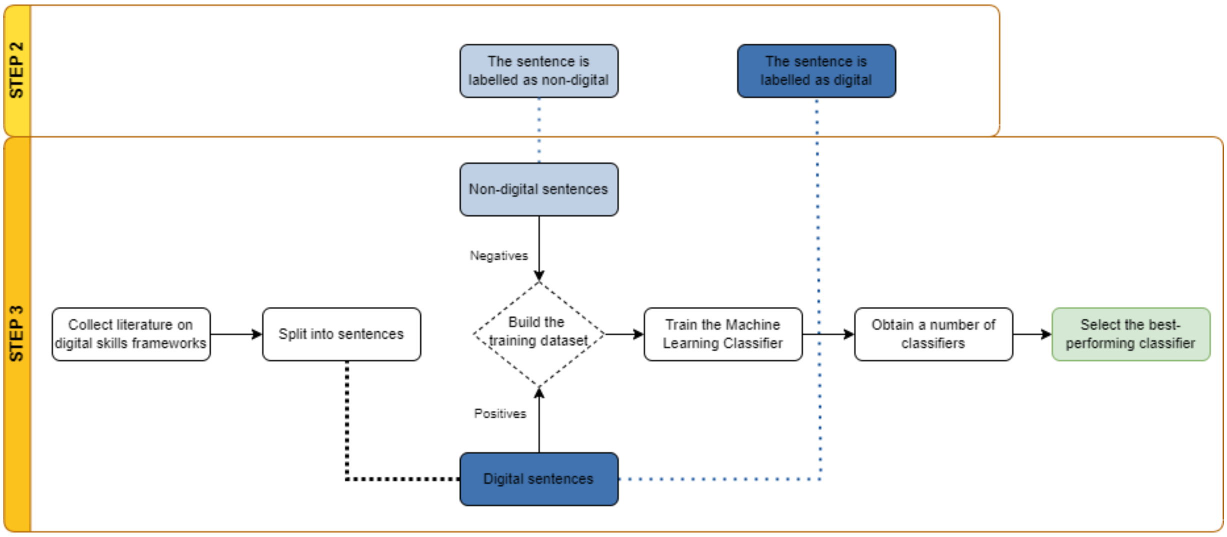


Figure 4: The third step to label digital ESCO skills and knowledge concepts

In the **fourth step**, the model selected from the third step is used to calculate the digital probability of ESCO concepts. ESCO skills and knowledge concepts are given as input to the classifier, which then assigns different scores based on title and description of the concepts. Based on a threshold which distinguishes digital and non-digital skills, the ESCO skills are then grouped into two output datasets (digital, non-digital).

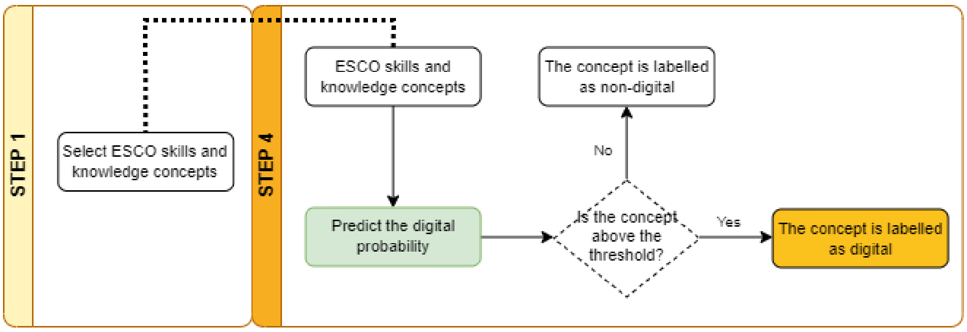


Figure 5: The fourth step to label digital ESCO skills and knowledge concepts

The **fifth** and final step is to compare the list of concepts *manually* labelled as digital (i.e. first step) with the concepts *automatically* classified as digital (i.e. third step). The final round of validation follows the following rules:

- If a concept is labelled as digital by the two approaches, it is automatically accepted as digital,
- If a concept is labelled as non-digital by the two approaches, it is automatically accepted as non-digital, and
- If a concept is labelled as digital by only one of the two approaches, it is revised and then labelled either as digital or non-digital. The revision is done manually by validators who must agree on all the labels, following the labelling rules defined in Annex 6.2 and employed for the manual validation (Step 1).

The result of this round of revision leads to the final list of digital and non-digital ESCO skills and knowledge concepts.

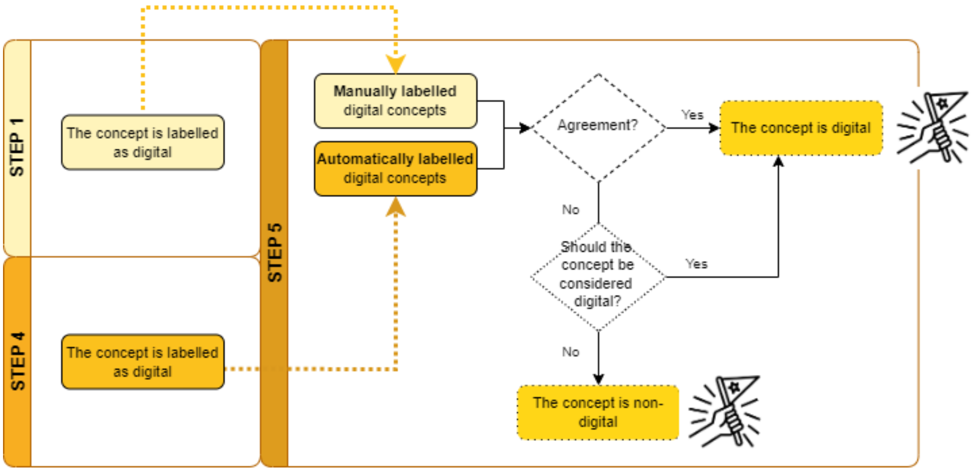


Figure 6: The fifth step to label digital ESCO skills and knowledge concepts

The following figure 7 provides an overview of the different sets of data employed in the steps described above.

Starting from the first step, the role of validators is to inspect all the ESCO skills and knowledge concepts, reducing the dataset of 13,890 concepts to 1,205 concepts which are the ones considered digital.

During the second step, once job vacancies are split and noise is removed, about 1,560 sentences are labelled as digital, in contrast with 7,608 labelled as non-digital. This demonstrates that about 40% of the sentences included in vacancies does not provide information concerning the skills required in the job. This value excludes the sentences removed by an automatic noise detector. Usually, such sentences include information concerning the hiring company, the contract benefits, and contact details. The nature of the digital skills from vacancies varies by sector, demanding for the use of common software tools in most cases, and for more advanced applications for the case of jobs related to the IT field.

In the third step the training dataset is constructed, summarising the digital sentences obtained in the previous step with those extracted from three main sources: the O*NET Technology Skills, the examples listed in the DigComp 2.2 publication, and the indicators developed in the European project Selfie For Teachers. As an output, different machine learning classifiers are resulting and compared on a validation set to select the best-performing model.

In the fourth step, all the ESCO skills and knowledge concepts are provided as input to the machine learning classifier. The Machine learning model assigns a probability to each ESCO skill and knowledge concept, indicating the likelihood for the concept to be considered as digital. A probability threshold is set at 0.96, resulting in a list of 1,958 concepts labelled as digital.

In the fifth step, the concepts labelled as digital in step 1 and step 4 are compared. The differences are compared manually by three validators and the final list of 1,201 digital skills and knowledge concepts is agreed.

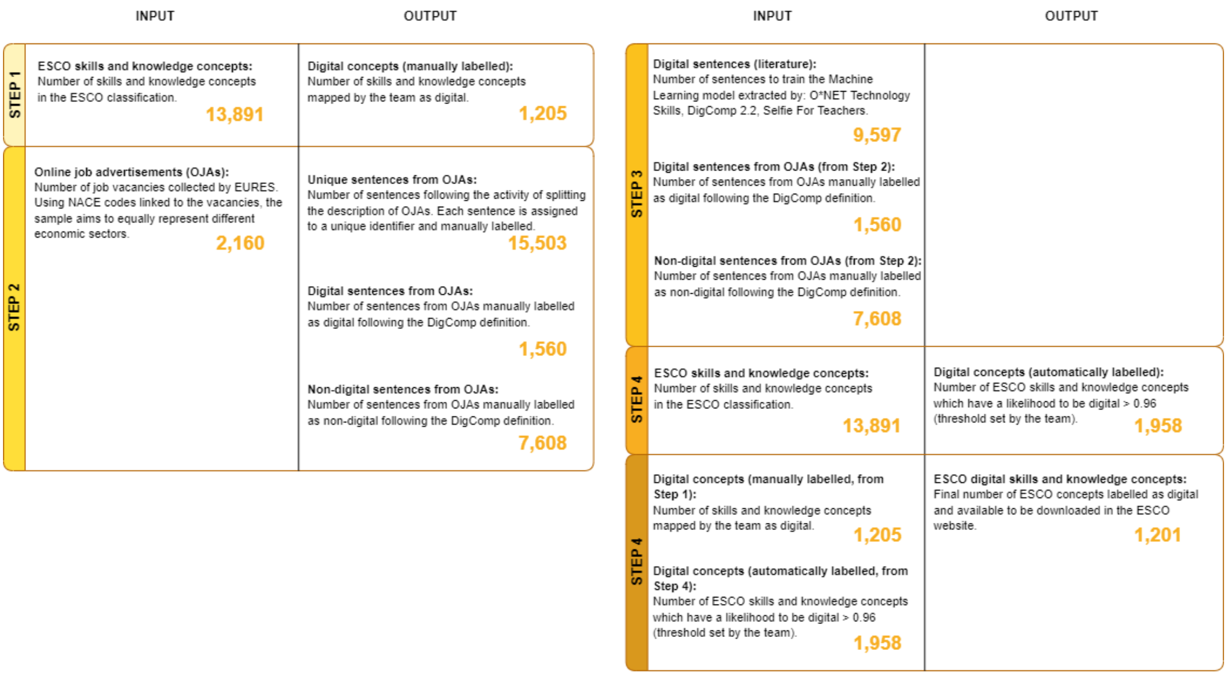


Figure 7: Input and output datasets for each step

3. Results

A total of 1,201 ESCO skills and knowledge concepts are labelled as digital. This includes: 718 skills (about 7% of all the ESCO skills, excluding the transversal ones), 475 knowledge concepts (about 16% of all the ESCO knowledge concepts, excluding languages), and 7 transversal skills (about 2% of all the ESCO transversal skills). The higher share of digital knowledge concepts is justified by the fact that ESCO includes a significant number of IT software and technologies. Looking at the reusability level of digital concepts, around half of knowledge and skills concepts is sector-specific, followed by cross-sector and occupation-specific concepts.

All the details concerning digital concepts are available in the ESCO portal, where citizens can navigate the classification, or download the datasets. This includes information on preferred terms, descriptions, non-preferred terms, relationship with occupations, reusability level, and more. Everything is translated in 28 languages (all the languages of European Member States, plus Icelandic, Norwegian, Arabic and Ukrainian).

The new list of digital concepts supports an analysis of the labour market with new insights, and the following charts provide a first understanding of the nature of digital skills in relation with the rest of the ESCO classification.

Chart 1 is composed of two pie charts representing the distribution of digital concepts in the first level of the skill hierarchy (on the left) and knowledge hierarchy (on the right). The hierarchies are not representative of different industries, but they rather focus on the nature of tasks and abilities required to perform duties related to an occupation. As expected, the higher share of digital concepts derives from groups focusing on computers and in general information and communication technologies. As for the skills hierarchy, groups on information skills, communication, collaboration and creativity, and working with machinery have about the same share. Examples of digital skills mapped in these groups are: collect data using GPS, edit digital moving images, maintain automated lighting equipment. Looking at the knowledge hierarchy, which is based on the International Standard Classification of Education (ISCED), following ICT concepts, knowledge concerning engineering, manufacturing and construction, and arts and humanities have the same share of digital concepts. Examples are: building systems monitoring technology, 3D printing process.

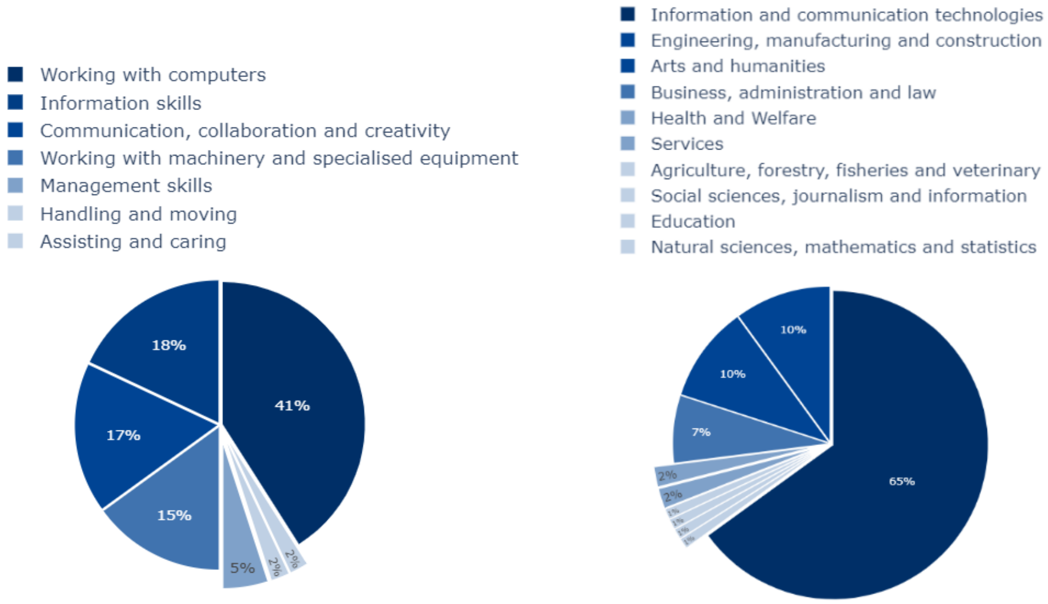


Chart 1: Distribution of digital concepts in the respective hierarchies

Having clarified the share of digital concepts from each hierarhical group among the whole dataset of digital skills, the following two charts look at the composition of such groups in terms of digital *versus* non-digital concepts. Charts 2 and 3 clarify that, except for the two groups dedicated to IT technologies and skills, the share of digital concepts is relatively low, reaching a top of 8% coverage for skills, and 19% for knowledge concepts. The order of groups by share is slightly shifted compared to what is highlighted in Chart 1, as the focus now is on the structure of each group, rather than the raw number of digital concepts.

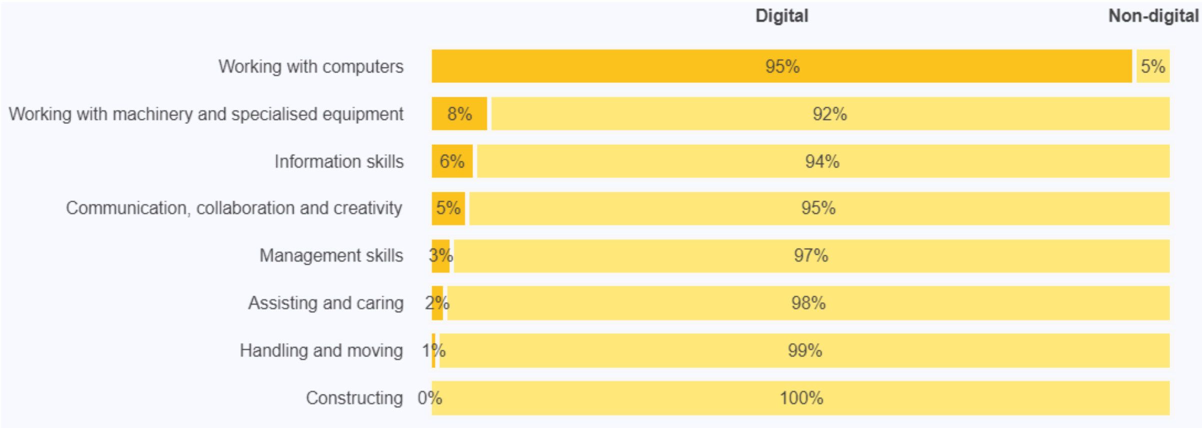


Chart 2: Share of digital versus non-digital concept in each hierarchical group (skills, first level)

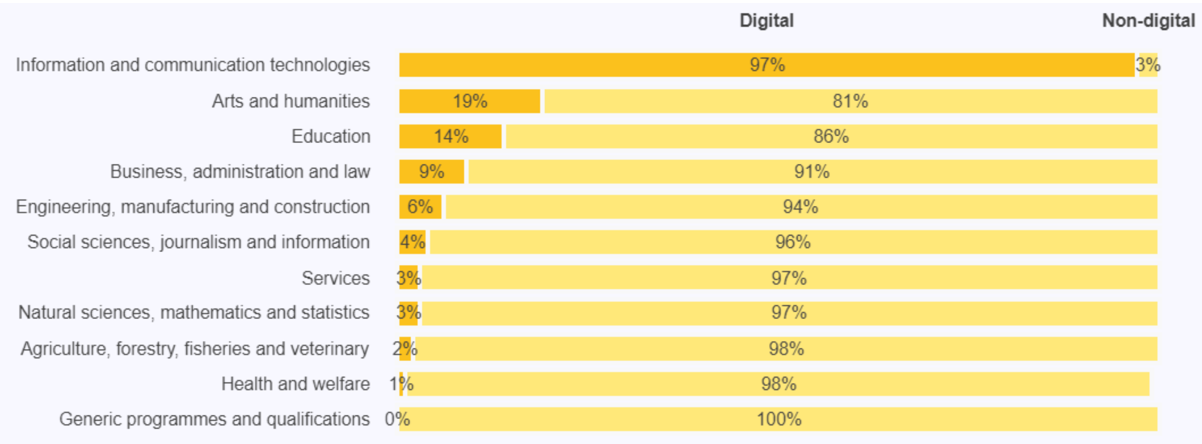


Chart 3: Share of digital versus non-digital concept in each hierarchical group (knowledge, first level)

One of the most relevant information from the ESCO classification is the link between occupations and skills. Chart 4 shows the share of occupations with at least one digital skill versus occupations not linked to any digital concept, among each group at the first level of the occupation hierarchy (ISCO first level). The calculation accounts for both essential and optional relationship. For every occupation group, except the one collecting elementary occupations, more than half of the occupations have at least one digital skill or knowledge concept linked.

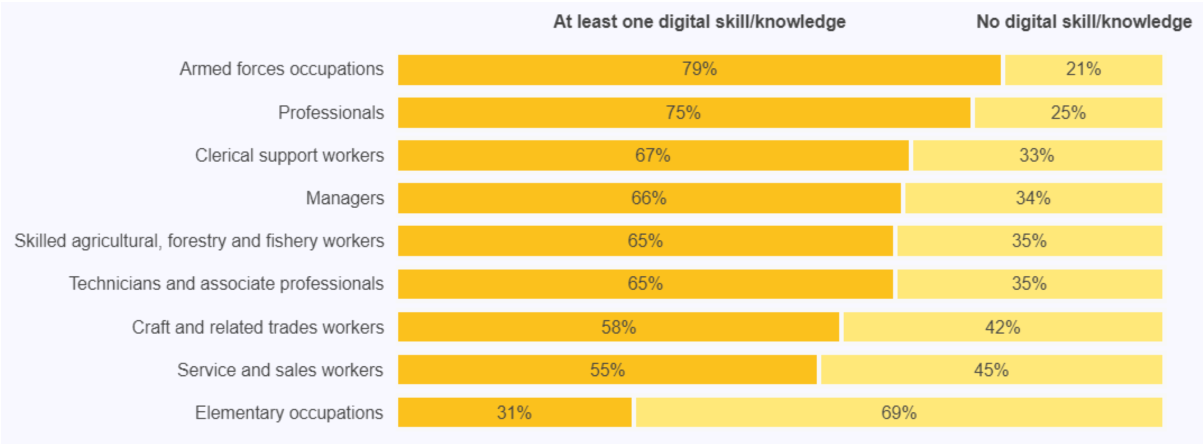
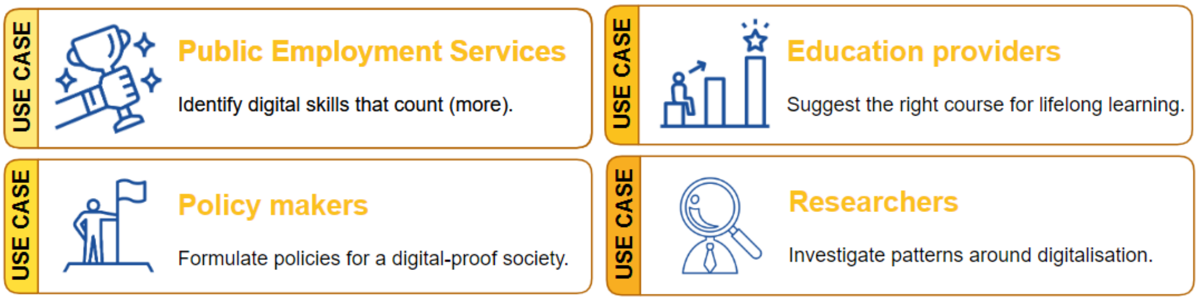


Chart 4: Share of occupations with at least one digital concept per ISCO level 1

4. Use Cases

The implementers of the ESCO classification compose a heterogeneous group of different stakeholders who participate in the development of one or multiple countries' labour market. Started with the purpose of supporting public employment services (PES) enhance labour mobility in the European Union, over the years ESCO has been adopted, among others, by private companies for their activities related to Human Resources, by education providers for a smooth integration of students in the labour market and to enhance life-long learning options, and by research institutes to analyse labour market trends and understand current and future patterns.

This section proposes examples on how to use the new digital labels for ESCO skills and knowledge concepts based on four main use cases: public employment services, education providers, policy makers, researchers.



Public employment services: identify digital skills that count (more). As part of the recurrent challenge of supporting job seekers matching with a suitable vacancy, PES officers need to minimize the skill gap between what is demanded by hiring companies and what is supplied by individuals. Today's occupations require the ability to navigate digital environments, and ESCO digital concepts can help pinpointing the skills needed in occupations – eventually focusing on a specific field of work. For example, among the group of handicraft workers, the most frequent digital concept is technical drawings, which is defined as *Drawing software and the various symbols, perspectives, units of measurement, notation systems, visual styles and page layouts used in technical drawings.*

Highlighting the relevant skillset to labour market actors would possibly reduce the unemployment period of jobseekers and searching period of employers.

Education providers: suggest the right courses for lifelong learning. An increasing number of individuals active in the labour market are interested in acquiring more knowledge, whether to increase their employability opportunities or for personal development reasons. While basic digital skills are now embedded within the process of learning or teaching (for example, using Massive open online course MOOCs or digital credentials), many are interested in improving their digital skills via courses. Using the ESCO classification, education providers can detect learning outcomes matching digital skills, and build recommendation systems that can target specific needs of learners.

Policy makers: formulate policies for a digital-proof society. Policy makers deal with innovations and interpret forecasts to support pathways towards an inclusive development –the Digital Compass is an example of such investments made by the European Commission. The list of ESCO digital concepts helps forecasting digital skills needed for the digital transformation of the labour market. Thanks to its advanced structure and international adoption, the ESCO classification can support the assignment of public investments based on a transparent and shared mechanisms.

Researchers: investigate patterns around digitalisation. Whether using official statistics (e.g. connecting ESCO skills to employment surveys via ISCO, as for the LFS), or continuous data (e.g. online job vacancies mapped to ESCO concepts, as for the Skills-OVATE project), the new list of digital skills and knowledge concepts allows for innovative research on labour economic and social topics. Researchers have the opportunity to compare digital *and* non-digital skills, looking for correlations with individual or industry indicators. Publications can be shared in a dedicated section of the ESCO portal.

5. Supporting Material

ESCO digital skills can be accessed via different channels.

***DigCompSkillsCollection* file**

As of September 2022, when downloading the folder with all the ESCO data from the Download Section of the portal, the file listing ESCO digital skills is included. It is important to distinguish this file, named *digitalSkillsCollection*, from the file of DigComp concepts, named *DigCompSkillsCollection*, which was first introduced upon the first release of the ESCO classification. The difference between the two documents is that the first represents the result of the labelling work presented in this report, while the second did not involve any work from the ESCO team and is entirely based on the categories defined by the DigComp framework. DigComp concepts are as well included among the ESCO digital skills.

In the *digitalSkillsCollection* file, each concept is identified by the Uniform Resource Identifier (URI), and presents information concerning preferred term, non-preferred terms, description, and other data. Information is provided for every of the ESCO languages.

***Skills* file**

For those who are interested in the full list of ESCO skills and knowledge concepts, this can be downloaded in the Download Section, in the same package as the one named above. We recommend using the latest version of the classification. Based on the file type, digital concepts can be filtered in the following ways:

- Using the CSV (Skills/competences, ≈ 9MB), green concepts can be filtered at the column *inScheme*, the concept type can be filtered at the column *skillType*,
- Using the TTL (ESCO v1.1.0 classification, Full RDF, ≈ 688MB), digital concepts can be filtered via the SKOS property *skos: inScheme*.

ESCO Portal, other sections

In the next months, the ESCO portal will be updated with the functionality to filter only digitalconcepts in the Skills/competences pillar.

More analytics on the digital concepts are available in a blog post that will be published in the Data Science and ESCO Section. This includes information concerning the methodology and interactive charts.

6. Annex

6.1 Machine Learning Classifier

This section provides details about the different steps that resulted in the final machine learning labeled set of ESCO skills.

The steps undertaken are as follows:

1. **Split OJAs.** EURES online job advertisements were split in phrases using a custom-made multilingual sentence splitter model. This model is finetuned based on the bert-base-multilingual-cased pretrained model.
2. **Predict skills vs noise (first batch).** The splitted phrases from the online job advertisements are pushed through a multilingual binary classifier that predicts whether a phrase likely contains skills, tasks or duties versus information about salary, the company, application details, location, etc. This finetuned classifier is also based on the pretrained bert-base-multilingual-cased pretrained model and its main purpose is to reduce the irrelevant content as much as possible.
3. **Manually label (first batch).** At this stage a random subset of all the filtered phrases is humanly labeled according to the DigComp definition and potentially undetected noise is further excluded.
4. **Run performance metrics on classifiers (first batch).** The labeled subset of about 4,600 phrases is split in training and validation data in order to build a BERT-based classifier. The best performing model is selected based on comparing AUC values on validation data.
5. **Generate embeddings, compute metrics, and manually label (second batch).** In order to streamline the process of manually labeling the rest of the phrases, another BERT model (that was finetuned on labour market data) was used to generate embeddings for all phrases (including already labelled ones from step 3). An agglomerative clustering was used to cluster these embeddings. For each cluster a number of metrics was computed:
    a. Cluster ID,
    b. Cluster size,
    c. Cluster averaged digital skill probability (using the model from step 4),
    d. Cluster variance digital skill probability, an indicator for whether a skill from step 3 is marked as digital for the cluster,

e. Sentence digital probability.

One more column is then left for the manual validation, where the validator assigns value 1 if the sentence includes digital skills, 0 if the sentence includes non-digital skills, and leaves the cell empty if the sentence does not include a skill.

The table below is an extract of the validation tables.

Sentence	Cluster ID	Cluster Size	Cluster prob (avg)	Cluster prob (var)	Sentence prob	Manual
experience in the field of Autodesk mechanical, AutoCAD, Inventor, Vault, construction and new development	1388	6	0.8129	0.1334	0.8950	1
Motor vehicle maintenance, motor vehicle repairs	1168	5	0.0007	0.0000	0.0004	0
Instruction and support for course participants	44	8	0.2502	0.1866	0.0005	0
confident handling of MS office products (Word, Excel, PowerPoint)	1041	10	0.9988	0.0000	0.9989	1
basic knowledge of SQL or databases is an advantage	1853	4	0.9989	0.0000	0.9989	1
Motivated and friendly colleagues	450	10	0.0004	0.0000	0.0002	
Class BE driver's license	1929	16	0.0003	0.0000	0.0003	0

This approach enabled us to more efficiently continue manual labeling as we could more quickly go over coherent clusters, while spending more time on spotting clusters with higher uncertainty (i.e. variance) in terms of digital skill probability. This process allowed us to focus on the more relevant cases rather than spending more time on large clusters of non-digital skills.

6. **Run performance metrics on classifiers (second batch).** At this stage step 4 was repeated to train a model on all labeled online job advertisement phrases, resulting in the final classifier that was used to predict digital skill probabilities for all ESCO skills.

6.2 Labelling Rules

This section provides more information concerning the rules for the activity of manually labelling ESCO skills and knowledge concepts (Step 1 from the methodology).

The ESCO Secretariat applies the definition of digital competence provided by the DigComp framework. Although this definition allows to anchor the labelling activity to a widely recognised European standard, a preliminary testing conducted on a sample of ESCO skills raised some questions related to the exact scope of the labelling activity. The main problem encountered is that most of the situations described by ESCO skills are nowadays carried out using digital tools: skills such as *develop business plans*, *draft tender documentation*, *deal with air traffic issue*s implicitly require the use of digital tools, even if neither the preferred term or the description mentions it.

Including all cases where the use of digital tools is implicit or highly probable is possible. However, there is a risk of inflating the list of digital skills in ESCO by adding concepts with no immediate connection with the digital domain, thus making it difficult for ESCO implementers and stakeholders to understand the methodology behind the mapping.

For this reason, for the manual labelling activity the ESCO team decided to further specify the notion of *digital technologies* provided by the DigComp definition by setting two additional labelling rules. A skill is labelled as digital if it meets **at least one** of the following criteria (in some cases, one criterion may overlap with another one):

1. Skills that **explicitly mention the exclusive use of digital tools to achieve their goal**. Skills where digital tools are listed as one of the many (non-digital) means through which the activity can be implemented should NOT be considered as digital. Unlike under (2), skill's end goal does not have to be in the digital domain, because the employment of digital tools or knowledge as mentioned in the label and/or description sufficiently highlight digital competence for the skill to be considered digital.
    a. *Includes* advise on online dating, design brand's online communication plan
    Why yes? The digital tool must be used to achieve a non-ICT-related goal.
    b. *Excludes* coordinate advertising campaigns, monitor traffic flow
    Why not? The digital tool is one of many non-digital tools employed to achieve the goal.
2. Skills that capture activity where **the main goal or core of the activity lies in the digital domain**
    a. *Includes* troubleshoot website, maintain computer hardware
    Why yes? The core of the activity lies in the ICT domain.
    b. *Excludes* sell gaming software
    Why not? The core of the activity lies in the selling and purchasing domain.

7. Glossary

Description: a text field providing a short explanation of the meaning of the concept and how it should be understood. Most importantly, it clarifies its semantic boundaries *(European Skills/Competences, Qualifications and Occupations Classification, ESCO).*

Digital skills and knowledge concepts: competence which involves the confident, critical and responsible use of, and engagement with, digital technologies for learning, at work, and for participation in society. It includes information and data literacy, communication and collaboration, media literacy, digital content creation (including programming), safety (including digital well-being and competences related to cybersecurity), intellectual property related questions, problem solving and critical thinking. (Council Recommendation on Key Competences for Lifelong Learning, 22 May 2018, ST 9009 2018 INIT).

ESCO v1.1: ESCO 1.1 is the last major release of ESCO. The Commission released it at the beginning of 2022, after a three-year period in which Member States had to map or adopt an ESCO under the EURES Regulation.

Knowledge: the outcome of the assimilation of information through learning. Knowledge is the body of facts, principles, theories and practices that is related to a field of work or study *(European Qualification Framework, EQF).*

Non-preferred term (NPT): can be synonyms (words with similar or same meanings) but can also be spelling variants, declensions, abbreviations, etc. They are regularly used by the target group (jobseekers, employers, education institutions) to refer to concepts that are described in ESCO with the preferred term *(European Skills/Competences, Qualifications and Occupations Classification, ESCO).*

Occupation: a grouping of jobs involving similar tasks and which require a similar skills set *(European Skills/Competences, Qualifications and Occupations Classification, ESCO).*

Preferred term (PT): each concept within ESCO has a designated, unique preferred name per ESCO language. It is called the preferred term and can be a single-word term or a multi-word term. The preferred term is used to represent a concept in ESCO in a specific language. Out of a group of terms with similar meaning, the one that best represents the concept is chosen to be the preferred term *(European Skills/Competences, Qualifications and Occupations Classification, ESCO).*

Skills: ability to apply knowledge and use know-how to complete tasks and solve problems *(European Qualification Framework, EQF).*

Transversal skills: Learned and proven abilities which are commonly seen as necessary or valuable for effective action in virtually any kind of work, learning or life activity. They are “transversal” because they are not exclusively related to any particular context (job, occupation, academic discipline, occupational sector, group of occupational sectors, etc.) *(European Skills/Competences, Qualifications and Occupations Classification, ESCO).*

Uniform Resource Identifier (URI): Each occupation, knowledge, skill and competence in ESCO is identified by a string of characters that follows a specific syntax: the Unique Resource Identifier (URI). Each URI is unique over the web, allows data from different sources to link to it, and is persistent *(European Skills/Competences, Qualifications and Occupations Classification, ESCO)*.