Lightcast

European Data Methodology

The modern labour market is affected by several phenomena: new technologies, aging populations and migratory flows are changing the labour market with an impact and a speedpreviously unknown.

People involved in planning strategies and policies for the labour market need information and new knowledge to study and understand these changes. The use of traditional and conventional tools is no longer sufficient (for cost and quality of the results): job postings data and Lightcast aim to set up a tool dedicated to the labour market able to provide job-related information on the web. The use of this information in decision-making allows measuring of actual job demand, evaluating its evolution, reducing the time-to-market of analysis and decisions and enabling multidimensional analysis (type of occupation, skill, industry, etc.).

Europen data supplies:

- Daily gathered data and monthly updatedstatistics
- Real time monitoring of the most requested occupations in the labour market
- Detailed skills requested in the labour market by its main stakeholders (companies, working agencies, HR)
- Skills related to occupations
- Emerging occupations and new skills
- An actual measure of mismatch between labour market demands and offerings

Being aware of the labour market demand enables:

- Quick knowledge of new market trends
- Awareness of companies requests
- Labour market analysis services
- Availability of deep level analysis

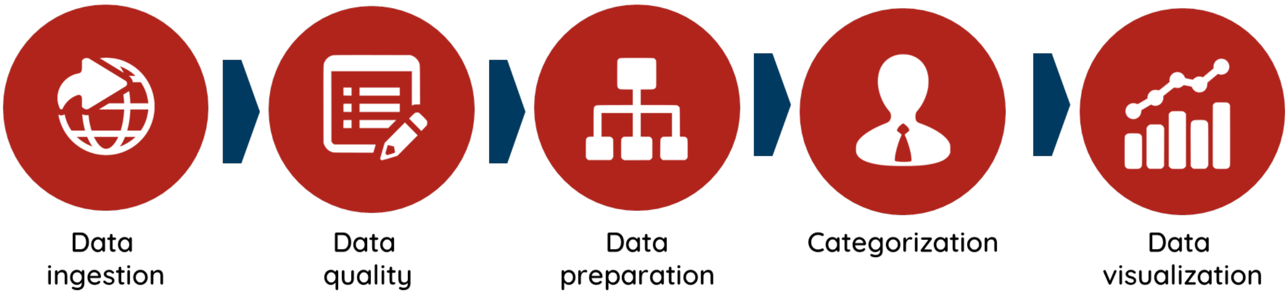
Lightcast EU dataset allows realizing statistical analysis on the evolution of the labour market vacancies published on the most important websites at various territorial levels: municipalities, provinces, regions or industrial districts.

Lightcast EU dataset is a statistical system and do not provide access to individual web vacancies.

The web vacancies posted on the web are expressed in semi or not structured texts and require, in order to extract information, a strict data processing methodology both from a technical and a scientific point of view.

Data processing procedures

The data quality and data processing techniques includes the following phases:

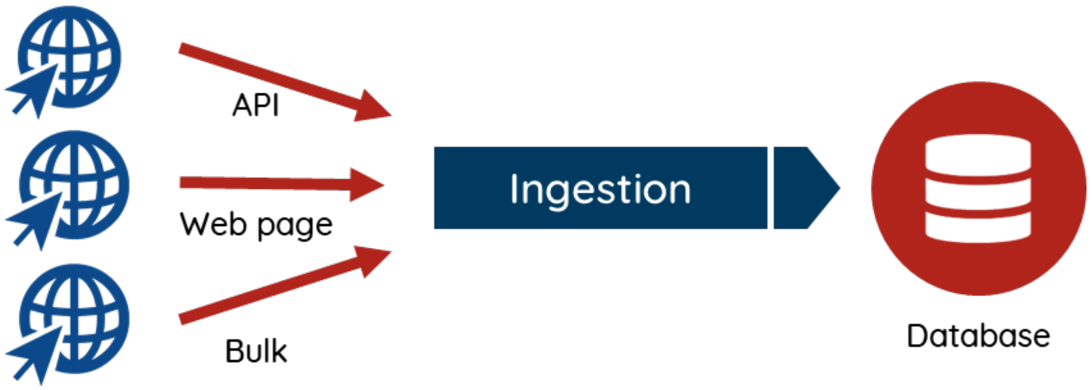


- Data ingestion: data gathering from the web sources previously selected and classified in order to guarantee their quality
- Data quality: data transformation and data cleaning techniques in order to guarantee data quality
- Data preparation: non structured data processing (vacancy’s title and description) to enable the following information retrieval phases
- Categorization: feature extraction and unsupervised learning techniques (machine learning) to extract, detect and link vacancy’s attributes (occupation, and skills) following international standards
- Data visualization: knowledge representation through the medium of visual paradigms following web usability principles

Data ingestion

Lightcast gathers web vacancies daily by means of various methodologies:

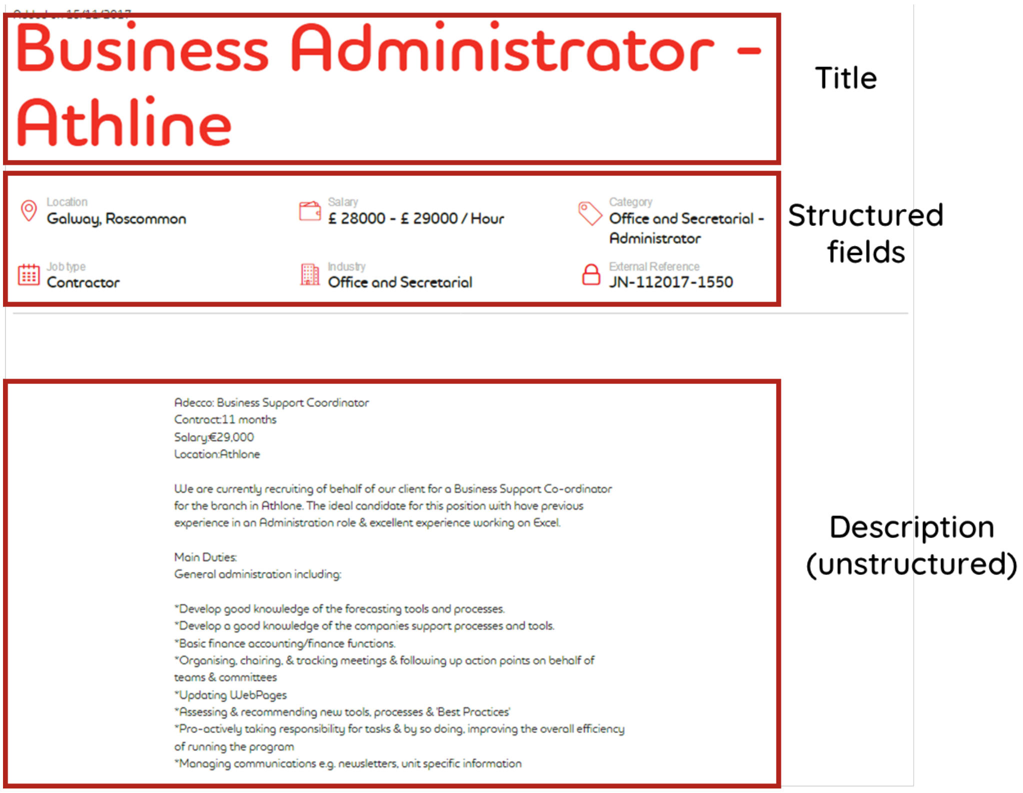
- Bulk extraction from data provided directly from the sources
- API interfaces
- Direct ingestion from the web pages (scraping)



A monitoring system allows controlling constantly data ingestion activities and their results in order to be able to quickly correct potential failures due to temporary unavailability of the web pages or to changes of data structures.

The structure of a web vacancy

A web vacancy can be published on the web in different ways or in different styles, depending on the source from which it comes from.



Added on: 15/11/2017

Business Administrator - Athline

Location Galway, Roscommon	Salary £ 28000 - £ 29000 / Hour	Category Office and Secretarial - Administrator
Job type Contractor	Industry Office and Secretarial	External Reference JN-112017-1550

Adecco: Business Support Coordinator
Contract:11 months
Salary:€29,000
Location:Athlone

We are currently recruiting of behalf of our client for a Business Support Co-ordinator for the branch in Athlone. The ideal candidate for this position with have previous experience in an Administration role & excellent experience working on Excel.

Main Duties:
General administration including:

*Develop good knowledge of the forecasting tools and processes.
*Develop a good knowledge of the companies support processes and tools.
*Basic finance accounting/finance functions.
*Organising, chairing, & tracking meetings & following up action points on behalf of teams & committees
*Updating WebPages
*Assessing & recommending new tools, processes & 'Best Practices'
*Pro-actively taking responsibility for tasks & by so doing, improving the overall efficiency of running the program
*Managing communications e.g. newsletters, unit specific information

However all vacancies are ascribable to a common data model including:

- Title
- One or more structured fields, defined by anaccurate format and taxonomy (publication date, industry, territory, etc.)
- A description in plain text (unstructured)

These attributes may vary in format and content from vacancy to vacancy and from source to source but they represent the informative basis all the following procedures rely on.

Structured fields can be considered strongly informative as, once detected, they already contains all the contents needed for their classification and to identify their domain.

Unstructured fields instead contain a lot of information but not exactly identified and they require additional processing and classification phases to be able to extract desired contents.

Data update

The systems collects data daily from the available web sources and its statistical interface is updated monthly, after the data processing and extraction phase.

Available sources

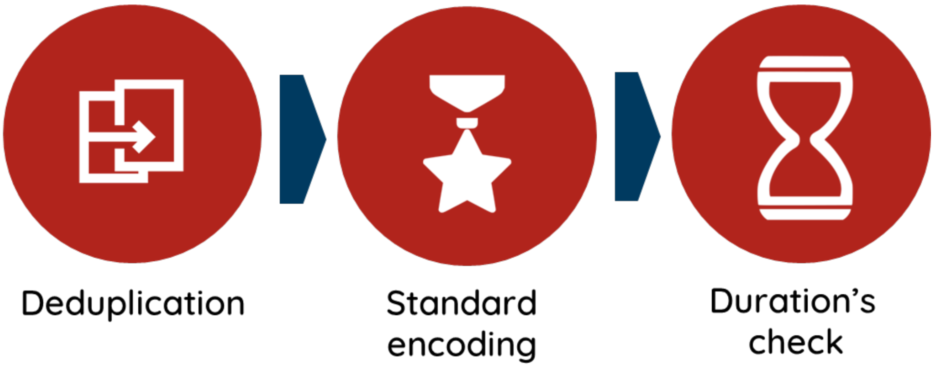
The sources inspected have been selected through a statistical model in order to guarantee completeness, reliability and data quality. The model takes into consideration the following parameters:

- Data completeness
- Data availability
- Update frequency
- Territorial coverage

The system currently gathers data from the following source’s typology:

- Web portals that publish Job offers
- Job offers aggregators
- Employment agencies
- Public employment services(PES)

Data quality

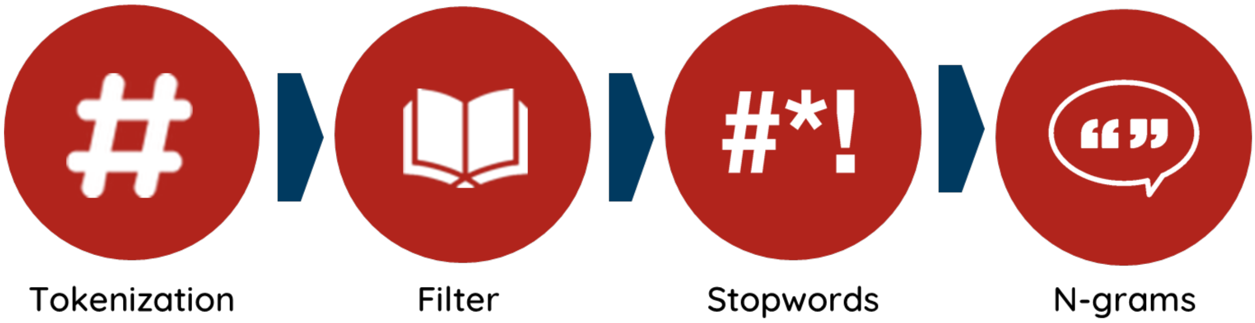


As stated above ingested vacancies contains both structured and not structured data (plain text). During the data quality phase, data are brought back to a standard data model by means of a strict methodology including the following steps:

- Deduplication of vacancies repeated on distinct sources or posted several times: text similarity techniques are used in order to identify potential duplicates and to trace them back to a unique instance; nonetheless similar vacancies posted more than a week after the closure of the previous instance are considered different
- Structured fields' encoding via international standards (NUT[1]- territory, NACE[2]- industry, ISCED[3]- qualification) or via specific taxonomies defined in order to support the language used by companies when posting a vacancy (i.e. contract). Structured fields already contains essential contents and all possible values (i.e. a drop down menu): mapping and matching techniques are user in order to bring back effective values to the standard taxonomy
- Duration's check: often vacancies are not closed on web sites after the right candidate has been identified; web sources independently carry out periodic remediation but it is not enough to guarantee data reliability. In this phase, further cleaning processes take place in order to identify a dynamic threshold (based on data distribution) after which a vacancy is automatically closed. Outliers here are defined as observations that fall above Q3 (3rd quartile) + 1.5 IQR (interquartile range).

[1] http://ec.europa.eu/eurostat/web/nuts
[2] http://ec.europa.eu/eurostat/statistics-explained/index.php/ Glossary:Statistical_classification_of_economic_activities_in_the_European_Community_(NACE)
[3] http://ec.europa.eu/eurostat/statistics-explained/index.php/ International_Standard_Classification_of_Education_(ISCED)

Data preparation



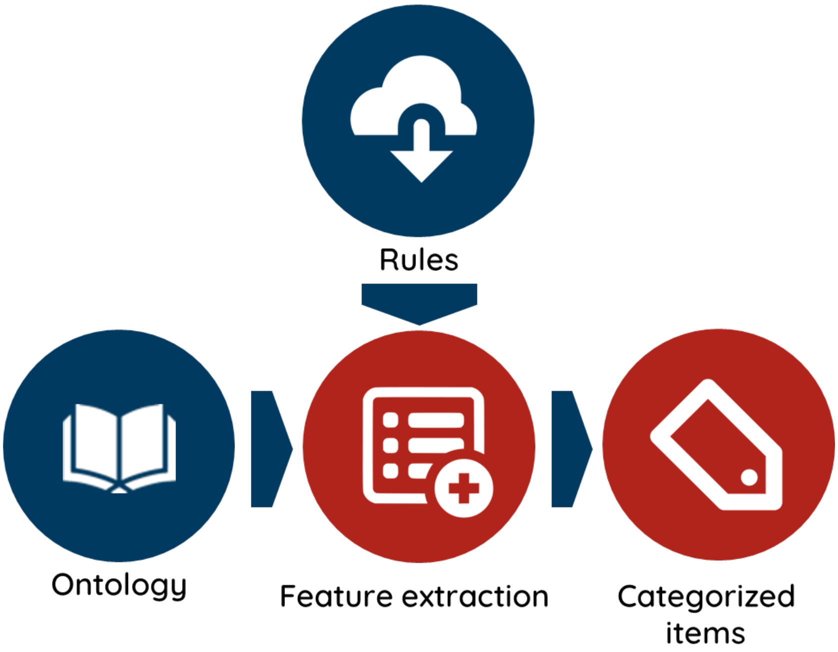
In order to be able to process the vacancies searching for useful information, plain text (unstructured data) undergoes some preliminary processing steps in order to prepare it for the following phases and to guarantee data quality and consistency:

- Tokenization: the plain text (token) is converted in a sequence of tokens (strings with an assigned and thus identified meaning), punctuation is removed and token can be later stemmed (the stem is the part of the word that is common to all its inflected variants)
- Filter: token are filtered by means of their overall frequency in vacancies, excluding common tokens (not significant) or rare tokens (not distinctive); token are filtered as well by means of their part of speech role (nouns and verbs are kept, adverbs, conjunctions, articles are discarded); finally short tokens are deleted paying attention not to discard significant acronyms
- Stopwords: token are filtered by means of a stopword list; this list includes common stopword dictionaries and domain specific word to be excluded (i.e. employed, company…)
- N-grams: tokens are combined in n-grams (n-gram is a contiguous sequence of *n* items from a given sequence of text or speech)

Categorization

Categorization allows the extraction of missing information directly from the plain text (description) of the vacancy by means of feature extraction techniques.

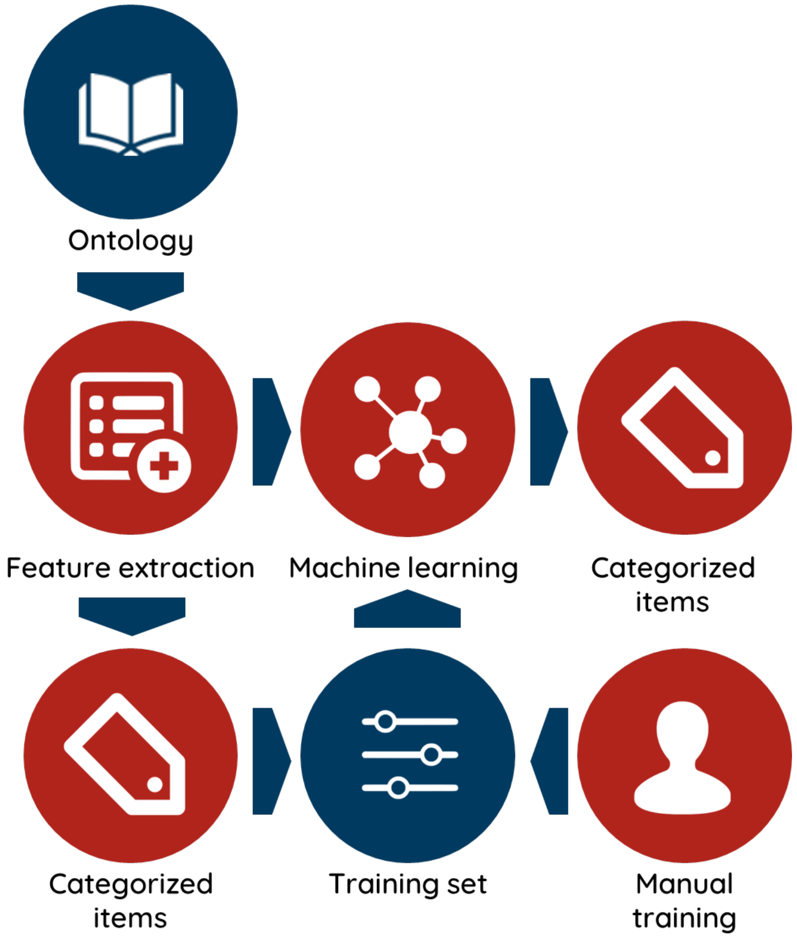
For all attributes (except occupation), the feature extraction process can be summarized as follows:



- Definition of ontologies or rules for the specific domain: for each attribute (skill, industry, …) dictionaries (ontology) or specific rules are identified in order to search the text for their instances; ontologies can be further empowered by means of synonyms
- Feature extraction: ontologies and rules are applied to text in order to find potential matches; similarity techniques are then applied in order to identify the most suitable occurrence

At the end of this phase each vacancy is linked to attributes on the basis of instances found inside the text; ontologies and rules may vary by means of the characteristics of the attribute; the more comprehensive are ontologies used the more efficient will be the result. In some cases, it may not be possible to identify suitable matches therefore the attribute is left blank.

Occupation’s categorization, due to its relevance in the system, undergoes a more complex processing phase in order to ensure better results.



The first phase, similar to the one previously described, outputs a group of categorized vacancies and a group of still uncategorized vacancies (usually about 10-15% of the total) for which feature extraction generated no results. In order to optimize results, categorized vacancies are included in a training set, later completed and validated manually, used to train a supervised learning algorithm (machine learning) to classify remaining vacancies. Acting like this the linkage between each vacancy and an occupation is assured.

Data visualization

Once ended the data processing and categorization phase, vacancies are arrangedin order to be analysed denormalizing data and implementing further filters (only vacancies concerning the first three groups of ESCO taxonomy, the territory and active during the last year). Now the dataset is ready to be published.

Information detail

For each vacancy, the following attributes are extracted and published:

- Publication date
- Expire date
- Occupation (4th level of ESCO taxonomy)
- Territory (NUT taxonomy) at municipality level where available
- Contract
- Educational level
- Industry (2nd level of NACE taxonomy)
- Required experience
- Working hours
- Source
- Skills (ESCO taxonomy)