Open Research Online

The Open University's repository of research publications and other research outputs

Structuring visual exploratory analysis of skill demand

Journal Item

How to cite:

Dadzie, A-S.; Sibarani, E.M.; Novalija, I. and Scerri, S. (2018). Structuring visual exploratory analysis of skill demand. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 49 pp. 51–70.

For guidance on citations see [FAQs](#).

© 2017 The Authors

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Version: Version of Record

Link(s) to article on publisher's website:

<http://dx.doi.org/doi:10.1016/j.websem.2017.12.004>

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data [policy](#) on reuse of materials please consult the policies page.

---

oro.open.ac.uk

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Structuring visual exploratory analysis of skill demand

A.-S. Dadzie^{a,*}, E.M. Sibarani^{b,c}, I. Novalija^d, S. Scerri^{b,e}^a KMi, The Open University, Milton Keynes, UK^b University of Bonn, Bonn, Germany^c Institut Teknologi Del, Toba Samosir, Indonesia^d Institut Jožef Stefan, Ljubljana, Slovenia^e Fraunhofer IAIS, Bonn, Germany

ARTICLE INFO

Article history:

Received 18 February 2017

Received in revised form 17 November 2017

Accepted 11 December 2017

Available online 21 December 2017

MSC:

00-01

99-00

Keywords:

Domain modeling

Knowledge discovery

Visual exploration

Ontology-guided visual analytics

Trend identification

Demand analysis

ABSTRACT

The analysis of increasingly large and diverse data for meaningful interpretation and question answering is handicapped by human cognitive limitations. Consequently, semi-automatic abstraction of complex data within structured information spaces becomes increasingly important, if its knowledge content is to support intuitive, exploratory discovery. Exploration of skill demand is an area where regularly updated, multi-dimensional data may be exploited to assess capability within the workforce to manage the demands of the modern, technology- and data-driven economy. The knowledge derived may be employed by skilled practitioners in defining career pathways, to identify where, when and how to update their skillsets in line with advancing technology and changing work demands. This same knowledge may also be used to identify the combination of skills essential in recruiting for new roles. To address the challenges inherent in exploring the complex, heterogeneous, dynamic data that feeds into such applications, we investigate the use of an ontology to guide structuring of the information space, to allow individuals and institutions to interactively explore and interpret the dynamic skill demand landscape for their specific needs. As a test case we consider the relatively new and highly dynamic field of *Data Science*, where insightful, exploratory data analysis and knowledge discovery are critical. We employ context-driven and task-centred scenarios to explore our research questions and guide iterative design, development and formative evaluation of our ontology-driven, visual exploratory discovery and analysis approach, to measure where it adds value to users' analytical activity. Our findings reinforce the potential in our approach, and point us to future paths to build on.

© 2017 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Exploratory discovery and detailed analysis of complex, unbounded scenarios are fundamental components of working and living in today's technology- and data-rich world. Obtaining a good understanding of such scenarios typically involves analysis of complex and/or large amounts of data. Such analysis often requires linking to other related datasets that provide context to fill in blanks in the core dataset(s) and aid understanding of the complete scenario. However, human cognitive limits make it easy to get lost in data as size, complexity and noise grow, hampering data exploration and exhaustive, in-depth analysis.

Effectively designed visualisation, by taking advantage of advanced human perception to recognise patterns and trends within

data, reduces cognitive load while aiding the user to obtain informative overviews and recognise and explore regions of interest (ROIs) in detail [1,2]. Ontologies provide a tool for structuring information spaces; in interactive visualisation this structure may be used to aid navigation through, querying and interpretation of underlying data content. Further, employing a uniform, underlying data structure supports representation from multiple, coordinated perspectives [3–5]. Ontology-guided visual analysis allows the strengths of each approach to complement the other, to enable intuitive yet structured, in-depth yet exhaustive analysis, increasing ability to obtain insight into complex data.

1.1. Motivation

The overall aim of this study is to determine the demand for jobs, and at a higher level of granularity, skills specific to a domain or industry sector and additional non-domain skills required to fill a job role. We aim to build intuitive representations of skill demand, that allow target end users to obtain a good understanding

* Corresponding author.

E-mail addresses: aba-sah.dadzie@open.ac.uk (A.-S. Dadzie), sibarani@iai.uni-bonn.de (E.M. Sibarani), inna.novalija@ijs.si (I. Novalija), scerri@iai.uni-bonn.de (S. Scerri).

of the demand landscape and how this bigger picture changes with location and time.

To meet our goals we scan job adverts using keywords corresponding to skills required to fulfil typical and specialised job roles in a domain. To allow us to evaluate our approach and also support in-depth analysis, we restrict our initial exploration to the *Data Science* domain and Europe. For completeness for this domain we capture a broad range of technical skills covering *data literacy*, *knowledge acquisition*, basic to advanced *exploration* and sophisticated *analysis*, and the *soft skills* needed to *communicate* the results of analysis within the technical domain and to a wider audience, using, e.g., *storytelling* and *data journalism*.

To ensure our analysis and results are reusable and extensible beyond this focus, we formalise our requirements for data capture, exploration and analysis using a set of ontologies describing skill demand and additional, relevant metadata, including target end user types (ranging from domain and technical experts to the interested but non-expert public); temporal attributes (dates and date ranges); and geographical location (at precision levels of country, city/town and latitude/longitude). We present the resulting knowledge framework as an upper level ontology that provides the backbone for structuring our analysis. This allows us to reuse the ontology for the more general case of skill demand analysis across any named domain and location. By capturing formally, also, the requirements of our target end users we obtain a framework that supports, further, the design and development of intuitive, interactive tools for task-driven, exploratory knowledge discovery and guided analysis.

To guide our study we explore the following data-driven research questions (RQs), with an aim to propose practicable solutions for the challenges that arise in the exploration of skill demand:

- RQ1:** what is the demand landscape for data scientists across the European Union (EU)?
- RQ2:** to what extent is demand matched by competence, as measured by skills in the existing workforce? Does this change when restricted to a specified location, time period, domain or industry sector?
- RQ3:** what support exists for a given individual – currently in the workforce or a trainee – seeking to identify and fill their “ideal” or desired job role?

Exploring these questions requires detailed analysis of job demand, using data from multiple, heterogeneous sources, with variation in attributes specified, terminology used and level of detail in job role descriptions. We employ use case scenarios describing our initial focus and target end user types to analyse the fuller set of challenges these RQs raise. Our findings feed into design and development of support for these, and ultimately, other end users exploring skill demand. For instance, successfully answering **RQ1** should allow us to provide intuitive support for, say, the general public, and therefore, a policy-maker at state level to retrieve information that answers the question:

How much movement would be seen within the EU for highly skilled statisticians seeking employment for the two-year period from mid 2016? Would access to more advanced, tertiary-level courses in object-oriented programming influence this migration?

Answering RQs 2 and 3 requires assessment also of existing expertise and the range of support in place for training and upskilling. This would allow a job seeker to answer questions such as:

What other skills would I need to find a job that allows me to combine my interest in journalism with the expertise I developed in my masters year internship building an iPad app to help pensioners keep track of their finances?

We will discuss the user- and task-centred process we follow toward answering such questions, and employ the working tools that support the exploratory discovery and sense-making tasks required to do so to review also the skill demand datastore we are building. This is a necessary precursor to designing intuitive support that will allow our varied target end users to independently explore and carry out further exploration and detailed analysis as needed.

Section 2 details one of a set of use cases developed to elicit end user requirements and illustrate the process we follow to generate an interactive landscape of skill demand. Section 3 reviews related work in job and skill demand analysis, and the value in visual, ontology-guided analysis for scenarios such as those we explore. We introduce our study methodology in Section 4, and discuss in Section 5 findings from our exploration of the initial dataset collected to describe skill demand. We feed these findings into a review of our initial requirements (Section 5.1), and formalise the outcomes in the *Skills and Recruitment Ontology* – SARO (Section 6). Section 7 describes the information extraction process that feeds into building the data store structured according to the requirements specified in SARO. We describe formative evaluation of our framework in Section 8 and illustrate, using the paths that participants followed to complete set tasks, tool design and functionality toward intuitive analysis of skill demand. We revisit our research questions and the use case scenarios in Section 9, to assess the extent to which our process enables target users to answer their questions. We also discuss, in this and within the context of each section, where future work will bring us closer to our overall goal. We conclude the paper in Section 10.

2. Use case scenarios

Recent technological trends show record volumes of data being generated and captured [6–8]. While this has the potential to contribute added value across the European economy and worldwide, efficiency of extraction and effectiveness in the use of its content impacts actual value recouped from this knowledge source. This modern challenge, due to and that can only be resolved through effective use of technology, was a key factor in the selection of our use case and initial application – to measure the gap between market demand for skilled analysts of *big data* and practitioners with the expertise necessary to meet this demand.

Our study methodology involves the exploration of skill demand and supply scenarios that consider the information-seeking and analysis requirements of the five target end user types identified in the early stages of the study [9], the: (1) policy- or decision-maker, (2) educator or trainer, (3) recruitment agent, (4) practitioner and the trainee or (skilled) job seeker and (5) the interested public. Based on the initial exploration of the requirements of our targets we created a task-based questionnaire that we use to guide our user-centred design (UCD) process, following an iterative cycle of design, development and evaluation. This paper extends the findings from this study [9] and reviews the initial requirements, to feed into the current exploratory and analytical stage. We focus here on the *practitioner* and the *trainee/job seeker* aspiring to this role – technical end users with data science-oriented backgrounds, and also non-technical users with good knowledge of the requirements for working with big data. Where relevant and for completeness, we highlight where requirements and our findings overlap with those of the other user types, as we explore in depth the data, and in an iterative UCD cycle, design for user-focused exploration and analysis tasks in line with users’ interim and ultimate information seeking goals.

Fig. 1. Process flow for the exploratory analysis used to extract and map skill demand. This extends the initial process flow described in Figure 1 in [9]. The approach taken in our study is guided by van Wijk's visual analytics model [2].

The [questionnaire](#),¹ which currently comprises four user tasks, has been updated to reflect feedback obtained from each evaluation session to date (detail in Section 8.1). We expand on these tasks to build a use case that describes support for the focus user type in this paper – the practising *technical data analyst*, often now described as a *data scientist*; this is a highly skilled individual with a set of core technical skills and varying capability in additional, related skills and competencies within the same and often across other skillsets. For completeness and to ensure utility and usability in the long run, we consider also the requirements of the informed, interested public, who may or may not have technical expertise. We build on these cases to illustrate the path(s) a user may follow to satisfy their information-seeking goals and measure the extent to which our approach meets our own and our target end users' requirements.

2.1. The technical data analyst or “data scientist”

Data Science is interpreted in a number of ways, due in part to the newness of the job title [7], and also the evolving infrastructure being built to cope with the very large scale, heterogeneous, dynamic data in today's increasingly advanced, technology-driven and dependent world. This has given rise to the need not just for technology experts but also other data literate employees who must work with a range of new tools, using technical and domain or sector-specific, as well as cross-disciplinary skills [10–13].

Differences in approach to managing the data deluge and defining what skills are needed to do so contribute to variation in the definition of a *Data Scientist* [14]. *Data Science* therefore presents a challenging, albeit interesting domain to explore, to identify core and desired skills and assess variation between the practitioner's definition of the field and that of the decision- or policy-maker who influences *upskilling* and the creation of new roles. A benefit in focusing on the practitioner is that over time these experts often also take on the role of managers or decision-makers.

This user (as all others) may, alternately, employ high level data overviews and a variety of specialised tools that support investigation of detail in ROIs, to serve varying information-seeking and exploration goals.

3. Related work

In line with our approach that uses the structure and knowledge provided by ontologies to guide visual knowledge discovery and analysis (see Fig. 1), we start with a look at work on automated and semi-supervised approaches to information extraction, focusing

on skill and job demand metadata. We then review studies of employment and skill demand data and scenarios, highlighting the formal models we extend for our study. We conclude with a look at the use of ontologies to model and guide the visual analytics process, to aid discovery of the knowledge contained within and support derivation of insight from the multi-dimensional datastore we employ.

3.1. Information extraction and enrichment

Job adverts constitute a useful information source for identifying and tracking qualification and skill requirements in the job market [15]. Online portals such as the dedicated job boards *Indeed*, *JobServe*, *Adzuna*, *jooble*, *XING*, *CareerJet*, *Trovit* and *CareerBuilder*² provide a useful mining source for state of the art text analysis approaches to automatically and continuously extract data [16].

Harper [17] carries out a study of the research methods utilised in investigating employment and skill demand in Library and Information Sciences. The aim is to recommend paths to follow to improve quality and coverage in data collection and the outcomes of analysis in such studies, and ultimately, reusability of findings beyond this scope. Recognising the utility of employment and skill demand data, effort has been employed in named entity recognition (NER) to address the gap between advertised job roles and workforce capability. Zhao et al. [15] extract and disambiguate terms (skill descriptions) from a large resumé database to build a taxonomy to aid the matching of skilled individuals to advertised job roles. Wowczko [16], who describes the use of text mining to extract and map skills listed in job adverts to defined occupations, notes that rapid changes in the modern world require greater automation for effective mining and tracking of key and evolving skills in an individual and society at large, to determine how such change impacts the economy. As in our study, Wowczko recognises the need for more clearly defined benchmarks for identifying skill gaps, to feed into the development of training resources for filling them.

3.2. Employment & skill demand analysis

Taking a cue from knowledge representation models created over the last two decades, we seek to formalise and enrich the information we capture on occupations, skills and the recruitment context, to optimise analysis and reuse in a variety of contexts.

¹ Evaluation documents and other supplementary material at: <https://doi.org/10.6084/m9.figshare.c.3929569>.

² <http://www.indeed.com>, <http://www.jobserve.com>, <https://www.adzuna.co.uk>, <https://jooble.org>, <https://www.xing.com>, <http://www.careerjet.com>, <http://job.trovit.com>, <http://corporate.trovit.com>, <http://www.careerbuilder.com>.

Table 1
Comparison of relevant existing models.

Concepts	Vocabularies				
	ESCO	Schema.org	e-CF	O*NET	LMI4All
Skill definition	•	•	•	•	•
Skill types	•	–	–	•	•
Occupation	•	•	–	•	•
Qualifications	•	•	–	–	–
Awarding body	•	–	–	–	–
Job posting attributes	–	•	–	–	–
Domain independent	•	•	–	•	•
Instance descriptions	•	–	–	•	•
Type	Ontology	Ontology	Model	Model	Model

Khobreh et al. [18] conduct a study within the human resources context that results in the ontology *Job-Know*. *Job-Know* links the knowledge, skills, and abilities taught in vocational education and training with corresponding competence and recognised needs within the labour market. Terblanche et al. [19] investigate employer demand based on online job adverts. They build a GATE-based [20] tool, *EDIT*, for *Employer Demand InTelligence*, that is guided by domain and process knowledge captured in the *Employer Demand Ontology (EDO)*. *EDIT* supports semi-automatic population of instance knowledge extracted from job adverts by matching to corresponding EDO concepts. Using adverts over a month in nursing and midwifery in Western Australia, collected from the job board *Sharing Environmental Education Knowledge (SEEK)*,³ *EDIT* was used to enhance and extend the knowledge base in EDO, allowing end users to determine employer demand by querying its SPARQL endpoints.

Other models that map closely to our topics of interest include the US Department of Labor's *Occupational Information Network project (O*NET)* [21]. This model aligns job titles and descriptions to skillsets. Its set-based job definitions, grouped into *six sub-skills*,⁴ contain information on hundreds of standardised and generic occupation-specific descriptors. Of these, one, under *technical skills – programming* – is directly relevant to our study.

Focusing on the EU labour market are the *European Skills, Competences, Qualifications and Occupations (ESCO)* ontology [22], the *European e-Competence Framework (e-CF)* [23] and the UK-based *Labour Market Information for All* database (*LMI4All*) [24]. The e-CF model focuses on competencies: “a demonstrated ability to apply knowledge, skills and attitudes to achieve observable results” in ICT. It is structured along four dimensions: business processes, e-Competences, proficiency levels for each e-Competence, and related knowledge and skills. *LMI4All* re-uses the O*NET descriptions of skills, abilities and interests, and further, describes other key labour market indicators such as future job openings, unemployment rates and vacancies.

We must provide, for our use case, (generic) *Skill* definitions, identify different classes (types) of *skills*, *Occupations* and *Qualifications*, and how they link to *Awarding Bodies* and *Job Postings*. To provide a single model that captures this domain knowledge we extend ESCO, Schema.org and other relevant work in SARO, the *Skills and Recruitment Ontology*.⁵ Table 1 looks at domain coverage in the five models and vocabularies that most closely overlap with our requirements. ESCO covers a large subset of these, bar the representation of actual job postings, which is defined by *Schema.org*. The three models O*NET, e-CF and LMI4All, while filling in other gaps, are not provided as re-usable vocabularies (i.e., as ontologies).

3.3. Visual, ontology-guided exploration & analysis

Exponential growth in data size and complexity and increased reliance on heterogeneous, third-party resources impact human ability to effectively extract, manage and make use of the knowledge content of the vast amounts of data that feed into task completion in today's data-rich and data-driven economy [16,25–27]. Visual analytics provides a valuable tool that complements advanced human perception and analytical ability with automated data processing and mining. This lowers human cognitive load in complex data analysis while improving the recognition of patterns and information hidden within data [1,2,4,28,29], support especially valuable in the analysis of complex, high-dimensional data.

Liu et al. [12] carry out a detailed survey of the different stages and components of high-dimensional visual analytics. Among others, they address the need for pre-processing due to the large number of dimensions and dataset size, to cluster data either along the dimensions defined or based on similarity in data point features. Liu et al., as do Hervás et al. [5], Kerren et al. [26], Turkay et al. [27], highlight the benefits in functionality for (automated) animation and interactive data exploration, that gradually reveal ROIs and trends in data, aiding the construction of understanding of complex, static and dynamic datasets. Inselberg [30], Heer et al. [25], among others, look at the benefits in multi-dimensional visualisation techniques such as scatterplot matrices and parallel coordinates. Kulyk et al. [31] examine the use of *focus+context* techniques to aid delving into ROIs in large, complex data.

Developing intuitive, user-centred, and therefore also task-centred, tools for effective visual analysis is not trivial, especially where a wide range of expertise, or lack thereof, is found in target end users [1,25,31]. Designing effective support for complex scenarios and data typically involves the exploration of the impact and use of multiple sub-components [4,12,26,27,29,32], following a process that may involve any or all of the following in an iterative cycle:

- data exploration often employing statistical analysis;
- preprocessing (including data transformation and/or reduction, and error handling); encoding and/or data (re)structuring;
- task-based and goal-oriented identification of optimal visual analysis techniques and tools;
- mapping of the whole or a sub-set of data to one or more visual representations;
- interactive exploration and contrasting of differing visual perspectives to identify patterns and reveal insight

Visualisation is being used increasingly in the Semantic Web (SW) and with SW technology, not just to present the results of analysis but as a driver for analysis and to enable more intuitive sense-making [5,33,34]. One tool proven to enhance the design and construction of powerful and intuitive visual analytics tools and guide effective use, for experts in visualisation and analytics, domain experts and other (non-domain and non-technical) users, is to build on the framework provided by ontologies describing

³ <https://www.seek.com.au>.

⁴ <https://www.onetonline.org/find/descriptor/browse/Skills>.

⁵ <http://vocol.iais.fraunhofer.de/saro>.

a domain and/or the analytical process [3,4,35]. Hervás et al. [5] formally model information spaces and the user's context, and map this to an information visualisation ontology, to support dynamic adaptation of visual representations of information to users' tasks and context. Khalili et al. [34] demonstrate potential for lowering effort to build user interfaces to support especially non-SW developers, by exploiting the structure of and metadata encoded in linked data. Elia et al. [33] use a taxonomy to encode elements in resumés as competencies, to aid comparison across differences in language and presentation in job role candidate selection. They use spider charts to visualise competency levels as named skills, along a set of criteria including experience and certification.

In the remainder of this paper we demonstrate, using SARO, the benefits in employing ontology-guided visual analytics for exploratory knowledge discovery and detailed analysis, to obtain a cycle of ever richer knowledge construction and application.

4. Analytical methodology

In order to answer our questions and those of our target users with a good degree of confidence we must let the data drive our exploration and analysis. Prior knowledge of the situation or context being modeled (or lack thereof) and the underlying data structure influence what approach is taken in sense-making and model construction. Because we rely on third-party data coming from multiple sources, automated mining and analysis approaches, while convenient and even necessary for large-scale data, are not a practical starting point. We extend van Wijk's visual analytics model [2] to define our methodology, illustrated in Fig. 1, that employs (interactive) visual analytics, relying on the human in the data exploration and analysis loop, to obtain an understanding, first, of data content [see also 1,26,29,31]. Working from this we iteratively build, verify and refine a model of our target users, their tasks and the resources required to work toward their end goals, as we widen the scope of our analysis to include new data.

We employ interactive visualisation as a working tool for intuitive, multi-perspective, exploratory knowledge discovery and detailed analysis of ROIs. An advantage in our iterative UCD cycle is an analytical process that in itself verifies both the tools and the skill demand and supply datasets we are building. We are therefore also able to reuse these tools to report the results of our analysis, and a selection for evaluation with different end users, to verify whether we are able to provide our target users with the information they require and intuitive tools to carry out further exploration.

The first steps in our process flow concern gathering and refining user requirements and building data overviews to feed into initial analysis. Our initial findings are reported in [9]; we follow on this in Section 5 with a discussion of the subsequent skill demand analysis, to provide a base to compare further results in our longitudinal study.

5. Early exploration

The human is a critical component in our visual analytics loop [12,27, among others] – to guide and refine automated analysis, interpret results and spot potential anomalies and (sources of) error by harnessing highly advanced human perception, intuition and experience. Initial user requirements were collected as part of the project proposal and an initial summary dataset collected based on them; this serves as the starting point of our analytical process. To ensure a useful guide for the user in selecting and using tools that optimise visual, exploratory discovery and detailed analysis, we iteratively review and update these requirements and the corresponding design.

The first dataset comprised summary data crawled from [Linked In](#), from 11 Mar to 05 Jul 2015, recording frequency of mention for a set of 46 pre-identified, core and specialised *skills*, grouped into 7 *skillsets*, required of data scientists for jobs across Europe. The data collection process, term identification and categorisation, carried out as part of the [European Data Science Academy \(EDSA\)](#)⁶ project, are detailed in [36]. The data was collected across three top-level dimensions: (1) skill, (2) geographical location and corresponding (majority or working) language and (3) time (daily). To eliminate complexity due to differences in translation of the terminology we focus in this section on the five English-speaking out of the 47 European countries in the dataset. It should be noted also that the data is skewed toward English and the United Kingdom (UK) especially – accounting for approximately 70% of term mentions in the dataset. However, relative frequency and patterns observed across the complete dataset are similar across all skillsets and languages.⁷ While we do not draw conclusions for all cases based on this subset, it provided a sufficiently representative sample for our initial human-driven, exploratory analysis. The complete dataset fed into a requirements review and subsequent data collection and analysis. We describe in [9] the initial exploratory analysis carried out using this summary data, but with a focus on trend analysis. We discuss here the skills analysis also carried out, to illustrate how these results feed into building our models.

The temporal component is a key feature in the data; we therefore drew basic timeline plots (see [9]) to obtain an overview and examine at the same time requirements for handling evolution of data characteristics and the data overall. These simple plots allowed also examination of detail in ROIs, to verify, for instance, skill co-occurrence and relative frequency of mention. Three key patterns were revealed: (1) significant dominance of one or two skills within each skillset, with ranges of up to a few thousand counts; (2) very low to no mention of up to three skills within each skillset; (3) a spike across all skillsets over a brief time span. Patterns (1) and (2) were repeated over the complete summary dataset. Further exploration using two other views on a complementary data snapshot (detailed in [9]) reaffirmed the trends observed in the timeline plots, and again in other subsequently explored, independently collected datasets (see Fig. 8, Sections 8.2 and 9.1). We discuss the implications of pattern (3) in Section 5.1.

To provide an alternative perspective on the (traditional) timeline we use an interactive node-link graph, built on the structure of the ontology described in [9] and Section 6, to reveal inter-relationships between skills based on mention in JobPostings over time. The graph is centred on the abstract SkillDemand concept (sub-class of Demand), with one node per day (114 over four months), and one node for each of the 46 skills per Demand object mentioned at least once on a given day – up to 5244 nodes per country (47 across Europe – up to 246,468 nodes). Fig. 2 shows the (initial) layout for mentions of the eight skills in the *visualisation* skillset in the UK. Colour-coding, filter and highlight options allow the identification of nodes of interest and immediately related nodes. However, the very dense graph, for what is only a sub-set of this summary dataset, hampers exploration and analysis.

To counter this and aid rather making sense of the data we exploit the upper level ontology summarised in [9]. By mapping the data structure to the ontology we are able to define templates that map to each concept; we focus here on the higher level concepts *demand*, *skill* and *time*. Each template object encapsulates object metadata and additional contextual detail, and the output is rendered based on the top-level resource type, direct relationships with other resources and the application context. We

⁶ European Data Science Academy: <http://edsa-project.eu>.

⁷ Additional, detailed snapshots can be found at <https://doi.org/10.6084/m9.figshare.c.3929569>.

Fig. 2. The node-link graph showing demand for skills in visualisation in the UK (~70% of the initial summary dataset), showing all nodes (instances of each concept) and the links between them. The result is a dense hairball with little support for knowledge discovery and sensemaking.

```

PREFIX so:<http://schema.org/>
PREFIX saro:<http://www.saro.org/saro#>

SELECT DISTINCT
  ?url ?date ?skillUri ?jobLocation ?skillDemand
WHERE {
  <skillDemandUri> rdf:type saro:SkillDemand .
  <skillDemandUri> schema:url ?url .
  <skillDemandUri> schema:datePosted ?date
  <skillDemandUri> schema:jobLocation ?jobLocation
  <skillDemandUri> saro:requiresSkill ?skillUri
  ?skillUri saro:frequencyOfMention ?skillDemand
}
ORDER BY DESC(?skillDemand) ?skillUri

```

Fig. 3. SPARQL template for SkillDemand, post data integration and preprocessing, simplified also for readability. This is run during user interaction to update the focus.

follow a methodology, detailed in [3], similar to that in Fresnel [37], that overlays visual *knowledge lenses* on data. Tominski et al. [38] apply a similar concept, using interactive visual lenses to adapt the presentation of detail in ROIs to suit user task requirements and context. Einsfeld et al. [4], Hervás et al. [5] follow a similar approach, mapping data to concepts in a backing ontology and relating these to visual metaphors that provide more intuitive representations of and the relationships within data.

We define templates that each comprise: (1) a SPARQL query that extracts the resource of interest and related data; (2) a corresponding visual template that encodes concepts and (directed) relations in the graph using size and colour, complemented by a detail (text-based) template providing more information about the ROI; the text detail templates are not discussed further here. Providing such *lenses* removes the need for processing and/or parsing by the end user, providing instead alternative visual perspectives on the data that suppress noise and reveal general trends and data characteristics of interest.

A parser (transparently) iterates over the RDF graph and runs the corresponding template query for each statement with a subject with matching *rdf:type*, for three *levels of detail* (LoD), basic,

(a) Focus on the skill *data stories*.

(b) Focus on the skill *d3js*, with tree depth increased by 1.

Fig. 4. Applying the templates to Fig. 2 reveals more clearly the Skill-SkillDemand relationships within the data. Focusing the graph first on *data stories* 4(a), then moving the focus to *d3js* 4(b) we see that the two skills never co-occur.

medium and full. Objects (resources or literals) matching a template are encoded and linked to or encapsulated within the subject node. The SPARQL template for the medium LoD for SkillDemand is shown in Fig. 3. Fig. 4 shows the application of the SkillDemand, datePosted and Skill visual templates to the summary data for the UK in Fig. 2.

To remove the clutter in Fig. 2 nodes are drawn for only the two focus concepts here – Skill and SkillDemand (over time). All other detail, such as JobLocation, derived from the JobPostings, is encapsulated within the corresponding template object. This detail is therefore visually suppressed in the overview, but may be retrieved on demand. Defining the templates using the ontology reveals additional insight, reflecting other work employing ontologies for visual knowledge discovery [4,5,35]. To focus on relationships between a Skill and other concepts it is related to a (compound) node is drawn for each instance of the SkillDemand concept, with default label DatePosted. A single node is then drawn for each (predefined) Skill, with a new (directional)

edge for each instance linked to another resource; in this case the *SkillDemand* instance node representing any date on which at least one job posting mentions the skill. Relative size of each *Skill* node is weighted by total *frequencyOfMention*, once per posting for the whole dataset. The spring layout therefore maps skill co-occurrence to distance between skills, and as a result also pushes outliers to the edge of the graph.

The user may restrict the number of levels drawn from the root (the abstract concept *SkillDemand*) of the current focus up to the maximum leaf depth. The user navigates through the graph by clicking on a node of interest to re-centre on that new focus. Fig. 4(a) shows the graph centred on the *SkillDemand* node for **21 Mar 2015**, drawing up to two levels from this focus. It therefore shows, navigating upwards: its parent (the root, abstract *SkillDemand* node); downwards: its descendants – *Skill* nodes mentioned on **21 Mar**; and at the same level all other *SkillDemand* nodes that mention each of these (child) skills. The mouse is hovered over the skill *data stories* (lower, left), highlighting it also in red. Further, this highlights its immediate neighbours in amber – all other *SkillDemand* nodes for the dates on which *data stories* is mentioned. We see also a cluster of skills in the centre of the graph (level 2 from the mouse focus); relative size shows that *data journalism*, which is also a bit further off centre, is required far less frequently than the other skills.

Moving the mouse focus across to the far right in Fig. 4(b), to demand on **04 Jul**, highlights its immediate neighbours: all *Skills* mentioned on that date. Keeping the mouse over the (04 Jul) demand node brings up a popup listing the five skills mentioned on that day and frequency of mention (across the UK). We see that *interaction* is mentioned significantly more frequently than all others on this date (see also Fig. 8, which shows similar relative frequency across this skillset, in a parallel coordinates plot for the data for all countries). While the **04 Jul** *SkillDemand* node can be seen in Fig. 4(a), its fifth child, the skill *d3js* is not visible because it falls outside the restriction of two levels from the focus, *SkillDemand* on **21 Mar**. All other children of the **04 Jul** *SkillDemand* node not also linked to that on **21 Mar** or one of its immediate children are also at a lower level, and therefore hidden. Maintaining the centre (at demand on 21st Mar) and increasing depth to three in Fig. 4(b) reveals the final skill, *d3js*. Its size indicates that, like *data stories*, *d3js* sees relatively low *frequencyOfMention*. With even lower mention than *data stories* and also co-occurrence with other skills, *d3js* is pushed to the outer edge of the graph. Both (relative) size and location confirm significantly lower mention of *d3js* compared to all others on **04 Jul** and across the complete dataset.

5.1. Requirements review

Further exploration using additional views on the summary data and a complementary data snapshot both reaffirmed the trends observed and prompted new questions, revealing further requirements for our use case. We illustrate in this paper additional challenges encountered during data exploration and analysis as data size and complexity increase, and work to iteratively refine requirements and design in our iterative UCD process.

Of the three key patterns we identified in the initial exploration exercise, pattern (3) highlighted a significant challenge in working with summary data. While ROIs can be seen the lack of context prevents further exploration to identify what lies behind them. From the user's viewpoint, this may lead to low confidence in analysis results, as it is difficult to verify the integrity or representativeness of the source data and the conclusions based on it. A key requirement therefore, highlighted in [9], is an information extraction (IE) process that results in detailed job postings, enriched as required with additional metadata; Section 7 describes the IE process built for this purpose.

Patterns (1) and (2) are repeated across the more detailed data – see Sections 8.2 and 9.1. Significant variation is seen in data density across two main attributes: (1) in the dominance of selected skills overall and within skillsets, (2) in matches for job postings by location. A key characteristic of the data collected to this point is, despite minor differences in skill distribution across datasets, a disproportionate number of postings for jobs in the UK, with a heavy skew toward Greater London and specific parts of London. Such localised density, repeated to a lower extent across the rest of the data, may be due to population density and/or clustering of the data and knowledge-driven industry in selected cities. Another contributor to data bloat is reposting of adverts by the same or multiple portals; ongoing work includes deduplication of such instances during IE (see Section 7). Future work will investigate other factors such as the impact of financial and industrial centres and (political) capitals of each country on data density.

At the other end of the spectrum is data loss during detailed analysis due to low granularity in location metadata, e.g., multiple branches of a company labeled at country level rather than locality or city. An associated challenge is verification of location during (automated) IE where ambiguous place names are in use, e.g., *East London* in South Africa annotated as *East* (part of) London, England. Future work involves the construction of (interactive) filters to allow users to flag, and link transparently back to improving IE, annotation errors such as these.

As the data grows in size and complexity additional functionality will be required to lower the resulting increase in cognitive load. This will include support for offline preprocessing and data aggregation, and online, interactive filtering, to maintain also interactive response especially during analysis of large data overviews. This must also align data type, user characteristics and tasks to the design of additional, alternative visual perspectives. Functionality for effective comparison of data elements with large differences in granularity and density will be key to revealing smaller maxima.

Our ultimate goal is to enable the identification of inter-relationships at multiple levels of detail, for individual skill demand and co-occurrence within and across job types and roles in general and within a domain or industry sector, as well as variation in demand across geographical location. This analysis must also compare skill demand against existing capability or expertise of *Practitioners* (skill supply).

Based on our findings to this point we detail in Section 6 the update to the ontology specified in [9], used to capture user, data and task requirements and guide data collection, exploration and analysis, and ultimately the presentation of the results of our analysis. We continue to feed these results into a cycle of ontology evolution, with the aim to continuously improve support for our own and our target users' information-seeking and analytical tasks.

6. Domain modeling

SARO, the *Skills and Recruitment Ontology*,⁵ provides a comprehensive representation of the knowledge required to define and correctly interpret job postings in the context of skills, competencies and the qualifications needed to fill a job role. Our modeling strategy consists of: (1) re-using existing knowledge in relevant standard vocabularies and ontologies (see Section 3.2); (2) following best practice in the development of new, open knowledge bases such that they can be linked to existing, related data. Fig. 5 provides a top-level view of SARO, which is centred around the following six core concepts:

Fig. 6. The information extraction and enrichment process.

saro:AwardingBody. an official or otherwise recognised institution certified to provide proof of the skills and competencies acquired in relation to a given standard following formal assessment.

SARO is complemented by a set of skill instances that are used to annotate job postings, and which comprise a key component of our skill demand analysis. We have previously referred to 46 (core data science) skills classified into 7 skillsets (see Section 5). We have since collected 739 skills with some overlap with this original set as part of the IE process. The skill instances are particularly important for the automated, ontology-based IE pipeline described in Section 7, as they are used to identify relevant skill mentions in job descriptions. This annotation task is not currently fully automated; the on-going ontology population task aims to capture also any additional relevant saro:Skill instances and metadata to feed into updating the saro:SkillSet categorisation, along with ranking of saro:Skills based on frequencyOfMention within a skillset and over the complete dataset. We aim ultimately to compare the categorisation obtained in this way with the views collected from practitioners in the field.

7. Information extraction and enrichment

The scale and diversity of our source data requires reliance on automatic detection and semi-supervised annotation of skills, to generate a knowledge base with sufficient breadth and scale to support in-depth analysis. We describe in this section the ontology-guided information extraction and enrichment (IEE) process employed, demonstrating also the impact of the process itself and the resulting knowledge store on our analysis and the validity of the results obtained.

The IEE process is based on the requirements identified for demand analysis (see Sections 5.1 and 6), encapsulated in SARO. The objective is to enable representations of skill demand meaningful both to the human user and that can feed into automated, algorithmically-dependent analysis, enabling each to complement the other to derive valuable insight and identify patterns and trends in the data. Fig. 6 illustrates how SARO is used to guide the IEE process carried out over multiple, heterogeneous data sources.

This involves two complementary pipelines: an ontology-based IE (OBIE) and a Wikifier-based¹⁰ pipeline (bottom and top, respectively, Fig. 6). Running the two pipelines in parallel increases

the total number of skills annotated (recall). Both further enrich the extracted data with geographical information, using concepts and instances from the *GeoNamesontology*¹¹ that correspond to locations (at levels city and country) specified in the job postings. Locations are extracted either based on precise latitude and longitude, where provided, or on the most specific location label.

The IEE process relies on SILK [40] to merge the results obtained from its two pipelines and reduce duplicate results, by: (1) aggregating annotation results for each JobPosting into a single entity, based on a previously specified unique identifier; (2) merging of duplicate skill instances identified by both pipelines. The process also supports manual review of newly identified skills, to feed into the ontology population task; this enables automatic annotation in subsequent executions of the OBIE pipeline.

The results are encoded in RDF, merging the (original) jobTitle, job description, postingDate and other relevant detail with the enriched semantic information comprising extracted, annotated skills and geographic metadata. RDF encoding provides two key advantages, both supporting reuse for further research and applications, by providing a self-describing, reusable datastore that supports our multi-perspective analysis approach (see Section 8); and promoting adherence to the principles of open publishing [41]. The formalised structure provided by RDF requires at most the use of wrappers for conversion for further processing in other tools and by third parties such as recruitment and education/training bodies.

Table 2 provides a summary of the contents of the RDF datastore as a rich, structured representation of skill demand in the data science domain across Europe for the period from Sep 2014 to Dec 2016. Sections 7.1 and 7.2 detail the two parallel pipelines.

7.1. The OBIE pipeline

This pipeline relies on various API connectors to the underlying data sources (including, as listed in Section 3, Indeed, jooble and XING²), configured to extract relevant data from each source in JSON or XML format. The raw data is transformed using the representation provided by SARO. The resulting data structures use concise representations that embody information extracted in a seamless manner, such as the job-specific skill and tool instances in SARO. This allows, for instance, recognised skills to be inherently

¹⁰ <http://wikifier.ijs.si>.

¹¹ <http://www.geonames.org/ontology/documentation.html>.

Table 2

Statistics for the ~850K annotated job postings in the RDF store as at Dec 2016, sorted by count per country, for the period Sep 2014–Dec 2016. The table is ordered top-down by total postings successfully loaded into the RDF store that subsequently passed the filter rules, based on the ontology, for input into the analysis tools. The double lines split the data at counts <100K, <10K, <500 and <50, respectively.

Country	No. of postings		
	Successful extraction from store	Distinct, named locations	At least 1 skill annotated
United Kingdom	159,842	139,771	103,112
Germany	144,757	126,566	103,848
the Netherlands	137,962	115,215	65,276
France	132,166	110,343	63,137
Ireland	31,246	25,283	20,093
Italy	22,875	15,131	9,090
Switzerland	22,261	12,587	13,072
Belgium	21,294	18,170	3,117
Poland	20,025	16,332	8,585
Spain	18,047	14,228	7,281
Portugal	15,056	8,208	6,685
Sweden	14,650	12,121	5,498
Austria	10,798	9,112	5,740
Denmark	10,491	3,414	3,009
Czech Republic	10,200	7,502	3,117
Romania	9,007	7,398	3,017
Hungary	8,507	6,852	1,796
Bulgaria	213	184	213
Norway	122	122	101
Malta	76	76	27
Greece	53	1	1
Slovakia	30	30	21
Liechtenstein	26	26	15
Gibraltar	19	19	19
San Marino	8	8	7
Finland	8	7	5
Estonia	7	5	5
Lithuania	7	5	5
Cyprus	2	2	2
Faroe Islands	2	1	1
Latvia	2	1	1
Andorra	1	1	1
Croatia	1	1	1
Luxembourg	1	1	1
Total	789,762	648,723	425,899

linked to known occupations (e.g., *SQL* and *PHP* to *Software Developer*).

Cross-posting of adverts in different job portals (and reposting in the same portal), requires deduplication of entries; we address this using SILK. A set of rules have been defined to determine similarity in extracted entities based on specified properties. Because *hiringOrganization* and *jobTitle* are consistent across data sources we prioritise these over other properties (e.g., *posting-Date*, *location*). A pre-defined threshold is then used to remove duplicate entries based on a weighted similarity score. The same rules are executed following the merging of the results from both pipelines, to remove any remaining duplicates.

The OBIE pipeline relies on the GATE framework [20], whose ontology-based module facilitates the identification of entities matching skill instances defined in SARO. GATE provides advantages over other NLP (Natural Language Processing) tools that would support OBIE, due to flexibility that allows its language processing functionality to be embedded into diverse IE applications. Our custom OBIE implementation works both alongside and with the Wikifier, by annotating extracted postings with matching skill instances harvested in the Wikifier pipeline.

7.2. The Wikifier pipeline

This pipeline complements the OBIE pipeline by executing the *JSIWikifierTool*¹⁰ to annotate entities based on a corresponding set of ranked pages from Wikipedia. An example of this *wikification*

process is shown in Fig. 7. This method has the advantage of recognising previously unknown, but relevant, skills based on collective intelligence provided by the community (Wikipedia), rather than on a pre-defined list as required for the OBIE pipeline.

Two additional sources are considered by this pipeline: *Adzuna* and *Trovit*². The data extracted is characterised by a number of important features: (1) multi-linguality, (2) representation in JSON format, (3) a cross-country view and (4) the presence of specific geographical and temporal components. Following the skill extraction process, the resulting annotations are aligned with SARO using label matching. If a matching skill instance is not found, the results are flagged for manual verification and on passing, added to the knowledge base.

7.3. Overall performance of the IEE process

Complete evaluation of the IEE process is currently in progress. As it impacts the validity of our analysis results we summarise here indications of performance, focusing on the information loss expected, given the integration of the independently developed components.

An evaluation of the OBIE pipeline's ability to annotate and extract relevant skills (in [42]) returned a strict F1-measure of 79%, which fares fairly well compared to results for human inter-annotator agreement (94%). Another on-going study compares the results of the JSI Wikifier to a manually annotated gold standard, achieving an F1-measure of 62%. The study in [43] investigating the

Fig. 7. Examples of wikification, using the JSI Wikifier for semi-supervised annotation of skills in a job posting, using the SARO framework. The corresponding JobPosting, converted to RDF, is shown on the right. Required attributes are shown in bold and the (predefined) skills extracted from the description field (with frequency of occurrence) annotated with a shaded background.

performance of SILK for interlinking entities, based on manually-provided linkage rules such as those coded for our IE process, reported F1-measure of up to 100%. This means that when following rigorous procedures based on an understanding of the entities being reconciled, using SILK following the parallel execution of our two pipelines is expected to have little impact on the F1-measure obtained for each pipeline. This should therefore not fall below 62%, with potential for a significantly higher upper bound; reliance on two complementary pipelines is bound to return higher combined recall than for each run separately. Work is in progress to measure precision and F-measures for the entire process.

8. Toward closing the skill gap in data science

A key element of our analysis is to determine skill correlation and ranking within skillsets overall, and in a named domain or industry sector. This is to guide, using SARO: (1) users in forming a profile of their capabilities, to map these to their “ideal” role or formulate a plan for updating their skills; (2) the definition of job roles and specific positions, to ensure that *essential* and *desired* skills are complementary and also correspond to relevant qualifications, capabilities and tools necessary to fill a role; (3) decision-and policy-making with regard to the allocation of resources for acquiring and updating skills and qualifications to fill recognised skill gaps.

To ensure SARO effectively supports both human-driven and automated reasoning and analysis, we must verify coverage and utility of the model for our target user types. In addition to IE we will ultimately also employ automated reasoning, both necessary for mining and analysis of big data. A detailed discussion of automated reasoning however falls outside the scope of this paper. We focus here on support for human reasoning as part of sensemaking, looking specifically at requirements for the practitioner and the trainee/job seeker. We use the evaluation tasks and use case scenarios to verify the extent to which we meet these needs and identify where further work is needed.

8.1. Evaluation methodology

As part of our user-centred approach we feed information from target end users into a cycle of requirements specification and wikification and (re)design, to ensure usability and utility, and practical adoption of our framework and the tools coupled to it [see, e.g., 4,31,32]. At this stage in our UCD cycle – iterating between design and development – we collect user feedback through formative, heuristic evaluation [44]. We use scenarios to walk users through the design and early prototypes [45] to assess potential in each tool for knowledge discovery and in-depth analysis. This approach

allows us to obtain detailed, qualitative information from target users throughout the iterative design and development process, and feed this back into improving and validating our analytical process and design [32,46]. Staggering evaluation over a long period ensures we move toward the *ground truth* and a more accurate and representative picture of skill demand as we continue to collect and explore data on the topic.

Feedback on data coverage and the initial overviews was collected during demos of early prototypes to policy-makers. This and a design review session covering the tools described in this paper, along with other related tools in a web-based dashboard, with a practising data scientist in the UK fed into a second cycle of user requirements review, (re)design and development.

Following this we carried out two heuristic evaluation sessions with two users each. The aim was to obtain additional qualitative feedback on the complete knowledge framework and also verify that our formal usability evaluation procedure as set up elicits the information we need to answer questions that pertain to our use case. To verify also data coverage and utility of the support provided for (visual) information retrieval and more direct querying of the underlying datastore, these sessions were targeted at domain experts – data science practitioners and non-scientists working in data science related functions.

Each session, lasting between one and two hours, started with a brief introduction to the project and the aim of our study. Participants then completed a demographic questionnaire. The flexibility afforded by informal evaluation allowed sessions to be tailored to participants’ backgrounds and availability; task assignment, from the set of pre-specified tasks, was mapped to participants’ backgrounds. An important aim was also to map usability, utility and preference of techniques to user types, visual literacy and tasks [46]. Following a brief overview of the different modules available for browsing and analysis, participants were encouraged to try each before selecting that they found best supported them in completing a task. Some sub-tasks were geared to specific tools and/or functionality, but participants were only required to use a specific tool where a task was designed to be answered using the tool in question (as in the case of Task 2).

A think-aloud protocol was used, and participants were encouraged to ask questions where necessary throughout the evaluation session. Each task was concluded with a task-specific questionnaire, and the complete session with an informal debrief, after which participants completed a more general post-evaluation questionnaire and the System Usability Scale (SUS) questionnaire [47]. The most recent version of our evaluation questionnaire and supplementary material are available online.⁷

All sessions were carried out remotely, with evaluator and participant sharing screens as required. This allowed participants in different locations across the EU to contribute differences in

perspective on expertise and demand due to location, language and local culture to the exercise, to feed also toward answering open questions assumed to be due to these differences. Two key are (1) choosing between thesauri that match local terminology and formal dictionary definitions; and (2) the gap between the picture of relative demand as relayed due to the English/UK skew, and participants' expectations based on knowledge of the local context.

The first evaluation session following the redesign, carried out in April 2016, involved two domain experts working in data and computer science, one each in France and Spain, and covered two activities. The dataset at this point contained ~300K postings across 22 European countries, with good coverage for 17 of these.

E1 Task 1. Job Role Creation: this required participants to describe their job role/function by creating a job advert to replace themselves, supported where necessary by information available from the datastore about job roles. The task ended with a tweet (up to 140 characters) summarising the advert; the aim being to elicit what each saw as core to their role.

E1 Task 2. Skills Analysis: participants were guided on a walk through the skill co-occurrence network (see, e.g., Figs. 4 and 12), to elicit additional information on skill-skill relationships using the skill-centric perspective on the demand data.

The second session, in Sep 2016, involved two researchers, one with a technical background and the other with a background in digital marketing, in France and the UK respectively. Both worked on projects in data science. This covered two main activities and was concluded with an informal review of a forms-based query tool. The dataset at this point contained ~500K postings across 33 European countries, with good coverage for 21.

E2 Task 1. Job Role Creation: as for Task 1 in the previous evaluation session.

E2 Task 2. Job-seeking Activity: this required participants to identify, from the datastore, roles that met their interests, but restricting their choices where relocation for an otherwise suitable job ruled it out.

These walkthroughs provide insight into the types of questions target users may seek answers to, and assess the representativeness and richness of the data from the end user's viewpoint, based on their experience and knowledge about demand in their locality. We use the different paths followed to complete these tasks to illustrate the influence of user feedback on our UCD process.

8.2. Visual, ontology-guided exploration & analysis

8.2.1. Task 1—job role creation activity:

We address first the tasks of measuring and ranking Skills and co-occurrence in job adverts. These tasks require examination of Skill-Skill and JobPosting-Skill relationships and, therefore, each concept and its attributes. JobPosting-Skill relationships can be derived from data content; whether a skill is required for a position or only desired (listed) can be determined by inspecting its description. Manual inspection is however feasible only for a relatively small sample, beyond which at least (semi) supervised processing is required [4,15,16,26]. Proximity of skills in job descriptions may be random, due to alphabetical listing and/or grouping by skillset; and may therefore not necessarily map to ranking or inter-relationships between skills (see, e.g., the job posting in Fig. 7). Determining (potential) relationships between skills therefore requires a greater degree of inference, even when done manually. As a first step we use skill (term) frequency and co-occurrence within each posting and the complete dataset to weight skills.

Support tools for completing Task 1 were two alternative perspectives on the demand data: a job-centric map view (not discussed further here; sample snapshots and a user guide may be found in the [onlinesupplements](#)⁷) and a skill-centric view employing parallel coordinates [30] (see, e.g., Figs. 8 and 9). This task requires exploration of the detailed JobPosting data extracted according to the knowledge framework embodied in SARO. To allow us to verify at the same time the utility of the visual analytics techniques employed and relate our findings back to the first set of views, we restrict our assessment to the same set of 46 Skills | 7 SkillSets (in Section 5 and [9,36]).

As part of pre-processing, therefore, we parse the input data to obtain the subset containing posts that mention at least one of these skills, which we analyse in addition to the complete dataset. The first batch of detailed data was extracted from job adverts posted on [GitHub](#)¹², [StackOverflow](#)¹³ and [Indeed](#)² between 13 Jul and 10 Nov 2015, using “Data Science” as a keyword filter. Out of the 12,632 matches found, 11,837 specify all JobPosting attributes required for our analysis (as specified in SARO—Section 6). We treat each skill as a dimension in the data, to allow comparison of metadata between skills, e.g., coOccurrence. Fig. 8(a) shows the use of parallel coordinates for visual display and comparison of all 46 skills simultaneously. Each polyline in the plot represents one of the 11,115 postings for which at least one of the 46 skills of interest is mentioned in the job description, intersecting each (skill) axis at frequency of mention for each posting. Job postings are “randomly” coloured to distinguish them, using a preset range of colours. Axes, with labels colour-coded by skillset, may be rearranged and shown/hidden to focus on a subset of interest [see, e.g., 12,30]. Combined with interactive query filters along each axis, visual AND queries may be formulated to track demand (as term frequency¹⁴) over time (datePosted, far left), to investigate Skill co-occurrence. The latter is illustrated in Figs. 8(b) and 8(c) for data stories and d3js.

Areas of high density for each axis and between adjacent axes indicate similarity for an attribute(s) across the dataset. Peaks and troughs, as outliers, often point to ROIs. For instance, in Fig. 8(a), discounting “data science” (the collection filter), “data quality” peaks at up to 22 mentions per posting — highlighted by the thick, broken maximum frequency trend line in magenta. Median frequency (broken green line in Fig. 8(b)) even for this relatively large dataset is 0 for all but “big data”, which has median 1 and maximum 18. Examining further distribution along this axis (in Fig. 8(a)) we observe that density increases with frequency, indicating a weighting toward more than 11 mentions per posting where found.

The need to manage increasingly large amounts of complex data is not an uncommon challenge in visual analytics [2,12,27]. However, screen real estate and resolution, processing power and human cognition place limits on how much data may be usefully displayed at once [1,28]. Beyond approximately 5000 postings interactivity in the web-based multi-dimensional parallel coordinates plot deteriorates significantly. For the filtered dataset containing just over 11,000 postings (Fig. 8(a)), implementing staggered load and redraw (that gradually build up the full picture) allows the visualisation to remain interactive. An additional benefit in the staggered approach is that the animation reveals patterns in the data that may be more difficult to identify in a static snapshot, especially in regions of very high density.

The datastore grew to approximately 300K postings in Apr 2016, nearly 500K by Sep 2016, and 850K by end 2016. As at

¹² <https://github.com/about/jobs>.

¹³ <http://careers.stackoverflow.com>.

¹⁴ We acknowledge demand may not simply map to raw counts, but must be qualified by, among others, desired and required skills — analysis considering these factors is in progress.

(a) Skill mention and co-occurrence for 11,115 job postings in Data Science; peaks, troughs and variation in density in this overview map to skill demand (frequency).

(b) View filtered down to three skillsets; and showing only postings mentioning each of *data stories* AND *d3js* at least once: **0** – all but the trend lines (thick, broken) are filtered out.

(c) Filter: postings mentioning *d3js* at least once: **9** polylines representing postings can now be seen. Note also the time axis shows *d3js* (nor *data stories*) not mentioned at all till early Oct.

Fig. 8. Interactive filters in a parallel coordinates plots used to explore skill demand and co-occurrence across job postings in Data Science from 13 Jul to 10 Nov 2015.

mid 2017 it contained over 2 million postings. Broken down further to visualise properties in each, among others, posting-Date, JobLocation, hiringOrganization, increases further data points, and therefore, complexity. Additional pre-processing is often necessary to support interactive visual analysis beyond what may appear to be a relatively low threshold, even using high-dimensional visualisation techniques, especially where data count far surpasses distribution across a large number of dimensions. Beyond a threshold of 20,000 we aggregate postings across spatio-temporal attributes, with links to further detail available by drilling down into ROIs. Fig. 9 plots the ~500K postings available during the second evaluation session. Each polyline, colour-coded based on JobLocation at country level, represents an aggregate per week, from the start of the capture period, containing postings that mention at least one of these skills in the job description, intersecting each (Skill) axis at total frequency of mention for the aggregate over the time period.

Participant feedback. All four participants listed broader and domain-specific technical skills as essential for the role they

proposed. Three, two of whom also manage others, included *soft (transversal)* skills: communication, leadership, negotiation, project management. The tweets (job summaries) contained a subset of the skills each participant listed – from selective/specific to relatively broad, and a job title/role summary. One included a link to more information, one a hashtag highlighting the most important skill for the role in question and one its location.

Both participants in the first session commented on the UK skew they recognised in the data; one therefore disregarded the job location sub-task, commenting that while the tool itself was useful the data skew meant the location-based distribution of search results was not reliable. The second noted the bias but chose to give a job location based on the results as shown. While the UK skew remained in the larger dataset used in the second evaluation, increased data coverage allowed the recognition of distinct patterns in other countries. The non-technical participant relied on the data while the other weighted responses based on knowledge of the local job market and expertise base. All participants, especially the two (one in each session) with lower experience found that

Fig. 9. A subset of ~500,000 postings across Europe listing at least one of 46 specified skills, aggregated by country in weekly periods. Aggregate counts for the *visualisation* skillset are overlaid on the plot (far right). The aggregate for *python* (far left) dwarfs all other skills at double the next highest – *big data*. We therefore lower the axis in the plot to an aggregate count per skill of 960 to reveal suppressed peaks for other skills – that *python* exceeds this limit can be seen by the corresponding aggregate lines extending above the plot boundary. This fell in the week recording the largest number of postings – 6741 in the UK, that starting 23 Oct 2015.

the visual skill categorisation aided ranking skill lists for their proposed role.

8.2.2. Task 2—skills analysis activity:

As discussed, determining Skill–Skill relationships, even manually, requires a greater degree of inference. This activity aimed to assess the support available using SARO to guide presentation of the data and navigation through the resulting information space, and therefore, support for tracking relationships within the data. We describe a sample journey with participants through the skill networks used in Section 5 to analyse the summary data. Fig. 4 shows the application of a simple template that uses aggregation to reduce density in the summary data, correlating this instead to relative size, to highlight relationships between features of interest. We extend this approach for the more detailed data (as in the sample posting in Fig. 7).

We narrow down the dataset first to focus on the top four countries by total posting count, in descending order: the United Kingdom (UK/GB), France (FR), Germany (DE) and the Netherlands (NL) (see Table 2). To continue to focus on the *visualisation* skillset and *d3* in particular, and include also the top two skills by mention overall – *databases* and *statistics*, we ran a second filter to extract only those postings where any of the following terms were found in the job description or annotated as a *requiredSkill*: *visual*, *d3*, *database* and *statistics*. Fig. 10 shows the query used to filter the datastore for matching URIs, using the corresponding geonames IDs.

We extend the lenses defined in Section 5 to cater for the increase in detail and data size. Additional visual cues are defined for the wider range of resource types, matching also colour-coding for outbound links from each node (type). The SPARQL template for the medium LoD for a *JobPosting* is shown in Fig. 12(a). The result set is then parsed to extract from the description field matches for the 46 pre-specified skills.

PREFIX schema: <<http://schema.org/>>
 PREFIX geo: <<http://www.geonames.org/ontology#>>
 PREFIX edsa: <<http://www.edsa-project.eu/edsa#>>
 PREFIX rdf: <<http://www.w3.org/1999/02/22-rdf-syntax-ns#>>

```
SELECT DISTINCT ?jobPostingUri
WHERE
{
  ?jobPostingUri schema:datePosted ?datePosted .
  ?jobPostingUri rdf:type edsa:JobPosting .

  ?jobPostingUri edsa:requiresSkill ?skillUri .
  FILTER ( regex(str(?skillUri), "visuald3databaselstatistics", "i") ||
           regex(str(?description), "visuald3databaselstatistics", "i") )

  ?jobPostingUri schema:description ?description .
  ?jobPostingUri edsa:Location ?geoLocationUri .
  ?geoLocationUri geo:parentCountry ?parentCountryUri

  VALUES ?parentCountryUri {
    <http://sws.geonames.org/2635167/> <http://sws.geonames.org/2635167/>
    <http://sws.geonames.org/2921044/> <http://sws.geonames.org/2921044/>
    <http://sws.geonames.org/2750405/> <http://sws.geonames.org/2750405/>
    <http://sws.geonames.org/3017382/> <http://sws.geonames.org/3017382/>
  }
}
ORDER BY DESC(?datePosted) ?jobPostingUri
```

Fig. 10. SPARQL query used to filter the dataset to the top four countries and job descriptions containing one or more of the terms “d3”, “statistics”, “visual” and “database”.

JobPosting, *datePosted*, *Skill* and *SkillSet* templates are used to build nodes for each instance of the corresponding resources. The visual template encodes node borders and out-going edges of *JobPostings* in purple, *datePosted* green, *Skill* blue and *SkillSet* olive. *JobPosting* nodes have default label *jobTitle*, and store, among others, *frequencyOfMention* of each skill listed as an attribute. *Skill* and *datePosted* nodes are sized (on a log scale) relative to *frequencyOfMention* and total number of postings on the date, respectively. Each *datePosted* node is linked to all jobs posted on that day, and each *JobPosting* is in turn linked to each skill mentioned in its description. *Skill* nodes are clustered around their (parent) *SkillSet*.

¹⁵ Total no. of date and skill matches is the sum after eliminating duplicates.

Fig. 12. Semantic graph for the detailed, enriched (job demand) dataset with templates applied to aid browsing and visual knowledge discovery. Selected node borders and all out-going links are colour-coded, and weighting (skill frequency, postings per date) mapped to node size.

Fig. 13. The two sequences show the relationships revealed as the user navigates through each graph and applies different filters, including both co-occurrence and exclusive skill mention (we draw an ellipsis to denote where non co-occurring skills would have been drawn), and how these change as the dataset grows – c.f. Figs. 4(a), 8(b) and 8(c).

task, less important information. Once drilled down to this more usable picture of the underlying relationships between skills, for a single posting or across the complete dataset, they were able to take advantage of support for navigating through and exploring the data.

A key requirement recognised was, therefore, to provide not only support for starting from a pre-specified focus, but also options to allow the user to restrict the initial view to display only content immediately relevant to this focus. The user may then proceed to navigate through the data to gain a better understanding of more distant ROIs. While starting with data overviews is the more common case, Heer et al. [25], for instance, demonstrate the alternative approach, that uses filters for data reduction and/or sampling in especially large datasets, to start exploration from an ROI. We have incorporated the feedback obtained into additional functionality for data pre-processing and interactive filtering. The graph is now rooted at first load on the abstract JobPosting concept with depth two. The initial view may also be set to suppress instances of any type. For example, hiding JobPosting instances removes the main cause of clutter, especially useful when focusing on Skill–Skill relationships for an ROI or as they change over time for a complete dataset (see, e.g., Fig. 13(a)(b)).

We now map, from Fig. 13(a)(a), an exploratory path as with the participants, starting with the graph zoomed in and re-centred on the skill *business intelligence*. The colour-coding aids navigation by revealing in the dense mass other skills it co-occurs with (blue borders). Moving the mouse focus to *big data* (top, left) highlights it (in red) and all postings in which both skills are mentioned (amber fill, red borders). The graph is then centred on the (abstract) JobPosting concept and its instances hidden (13(a)(b)). This reveals all skills mentioned in the postings in Germany, 29 out of the 46 predefined skills (~20K out of ~300K postings), grouped into distinct skillsets (encircled to aid identification). A time filter highlights with a pink (fainter) border the skills mentioned in Oct 2015. The following snapshot (13(a)(c)) moves the focus to *statistics*; this is one of the skills required in the job posting in Figs. 7 and 12. All other skills required for that posting, including *visualisation–interaction* and *machine learning*, remain in the view while non co-occurring skills in other skillsets no longer appear – the position of each is manually marked with an ellipsis. Finally, the focus is moved to *data mining* (13(a)(d)), the second most frequently mentioned skill overall; this in turn updates the skills in the view to match co-occurrence for all postings in Germany.

Fig. 13(b) plots a similar path for Germany for the dataset which extends to end 2016. At ~103K (out of ~850K), it is five times the size of the base for Fig. 13(a). Two filters are applied: first the SPARQL filter in Fig. 10; Table 3 shows this reduces the dataset to 29,862 postings. The second filter is run as the graph is built, retaining only those postings for which at least one skill in the *visualisation* and *maths & statistics* skillsets is mentioned, reducing the dataset further to 2959 postings, listing together 27 of the 46 pre-specified skills, on 197 dates starting on 03 Feb 2015. Additional time filters show data distribution for the resultset to be sparse till Jun 2016, when it starts to increase gradually through to 17 Dec 2016.

8.2.3. Task 3—job-seeking activity:

This task required participants to explicitly consider location in searching for matching skills (in postings). Two tools were used – a map-based view and a forms-based, guided query tool. The map visualisation focuses on location; therefore while we apply all relevant feedback across the suite of tools being developed in the larger study, to retain our focus on skills we do not discuss it further here.

A significant challenge we face, especially as data size grows, is obtaining useful overviews and narrowing down to ROIs, while

retaining access to the detail in the surrounding context and other more distant ROIs. We currently utilise a mixture of interactive visual/widget filters and simple text search in the visualisation tools discussed in this paper.

The forms-based data query/browser, while not a visualisation tool, provided us with feedback for implementing additional filters to support the visual analysis. The data browser provides two views, narrowing down from:

- **location** (*country to city*) to **skill**, or
- **skill to location** (*country to city*).

Participant feedback. The tool was described by one participant as utilising a “funnel” effect that eliminated information overload: starting from a single perspective or attribute and selectively providing additional filter options to guide the user toward a final result, while “funnelling” out noise. Both participants expressed a preference for the provision of a filter of this sort at the start of the exploration process. Among the reasons they gave two are key – which, incidentally, led to the development of the prototype: (1) to provide a preview of the structure and content of the dataset, (2) to reduce data density and narrow down quickly to an ROI. Both participants also used the analogy of searching for a new home on an estate agent’s web site, which typically provide a set of basic search filters, followed by more sophisticated tools for deeper exploration of (narrower) ROIs.

On-going work is addressing three additional key functions of this data browser, building on the ontology structure to: (1) provide an intermediary between the independently built analytical tools with complementary functionality, but which remain to be coupled; (2) enable end users to construct simple, visual, statistical summaries of data content; (3) enable download of data subsets and the results of analysis for reuse in third party tools.

We review, in Section 9, insight obtained through the use of our framework for user- and task-centred visual exploratory discovery and analysis, to feed these and lessons learnt from interaction with end users into proposals for further (re)design.

9. Discussion & lessons learnt

Our study aims to build visual analysis support for our target end users and to provide working tools that augment also our own capability for investigating the skill demand landscape. To measure progress toward our goals we revisit our research questions, to probe more deeply the insight gained through evaluating our approach and its accompanying design and tools.

9.1. Verification of visual, ontology-guided model

We have discovered patterns that have persisted through independently collected and successively richer datasets. We revisit these with a focus on cases where we obtained unexpected, new, or seemingly contradictory or anomalous results, and discuss what insight each led to.

RQ2. looks at representativeness of our dataset and results, overall and for specified locations and periods. A challenge we identified early in the study was a skew toward posts in the English-speaking countries and especially the UK, due to differences in terminology across the EU, regardless of language, in addition to gaps between formal translation and actual terminology usage. Resolving this is key to confidence in decision-making based on the contents of the datastore and analysis results, and therefore adoption of our approach. An important comment during early evaluation was that while added value in the framework was clear, there was still high potential for unreliability of the results because

of the skew in the data, a situation especially pertinent to end users working outside the UK.

We build on the framework provided by SARO to track skill demand trends across time and location. Looking again at Fig. 11 and Fig. 13, the former draws the skill network for the top four countries by skill mention – the UK, Germany, the Netherlands and France, while the latter plots a series of snapshots for Germany alone to contrast patterns seen at two fixed points in our data collection process. The increase in coverage across the skillsets for the two datasets in the sequences in Fig. 13 is not remarkable, even with the pre-filter applied to the larger dataset. In addition to 13(b) spanning a longer period we are gradually improving precision and recall in our IE process; we cannot conclude that this necessarily represents increased demand overall. We examine more closely, rather, relative change in data density, the absence of previously persistent patterns and large or seemingly anomalous changes, as well as the appearance of new patterns.

Looking at Fig. 13, the most noticeable pattern that has persisted (through our analysis) is the dominance of *statistics* and *databases*. Noticeably lower relative count for *databases* in 13(b) is accounted for by the second filter, which retained only those postings mentioning skills in the two focus skillsets. Of the eight skills in the *visualisation* skillset (cluster on far left, 13(b)(a)), *data stories* and *data journalism* are missing. In 13(a)(b), bottom, by contrast, we see only half of the skillset – *tableau*, *interaction*, *d3js* and *infographics*. Of the eight in *maths & statistics* (top, right, 13(b)(a)), *calculus* and *rstudio* are missing; in 13(a)(b) *rstudio* is seen, but *probability*, *calculus* and *linear algebra* are missing. In addition to the two top skills overall, *python*¹⁶ (advanced computing), *big data* (general), *business intelligence* (domain expertise), and to a lesser extent, *tableau*¹⁷ (visualisation), record higher than average coOccurrence and frequencyOfMention in both datasets. *Python* dominated especially the UK data in the last two datasets generated. We however saw relatively few counts for *rstudio*, despite anecdotal evidence indicating popularity close to that for *python* for big data analysis. This may be due to use of the IDE as a keyword rather than the statistical programming language's name, *R*.¹⁸ We will in future work adapt our filters to include synonyms for *R*, bearing in mind challenges associated with obtaining true positives for cases such as this single character, case-agnostic word.

Importantly, our results are aligned with discussions in the literature on the skills seen as essential in the emerging and highly dynamic field of data science. Recent literature on the evolution of the data scientist typically highlight overlapping sets of skills and tools for or commonly used in complex data analysis. Among the most frequently cited as vital for this emerging role is *statistics* [7,10,11,13] – the second highest skill by frequency of mention in our datasets. Ability to capture and manage large data stores or *databases* – the most frequently mentioned skill in our store – and carry out sophisticated querying across multiple, distributed stores also rank high [12, among others]. Specific tools and products, notable including *NoSQL*¹⁹ and *Hadoop*,²⁰ respectively, are often listed as desired or even required for roles that include data capture and querying from dynamic, heterogeneous sources. Experience in presenting data overviews and the results of complex, often long-ranging and dynamic analysis in more intuitive visual form, using off-the-shelf tools such as *tableau*, is another highly desired skill. Visualisation packages such as *D3js*²¹ allow more flexibility in building custom, web-based, static or interactive visualisations,

but also require a good degree of *programming* capability. Both “skills” (products) see relatively frequent mention in our collection of job postings.

We see here non-trivial relationships between skills. An ongoing task is (re)categorisation of the original 46 and the additional (739) skills identified. This task must take into account skills, including experience or capability in the use of named tools and products, that cross skillset boundaries, and the identification of synonyms – features that will provide additional context for enriching SARO.

Semi-supervised annotation, aided by functionality for multi-lingual term identification and annotation in the Wikifier, complement the human-driven annotation and categorisation tasks. Analysis results feed back into increasing coverage and accuracy during ontology population, in a cycle of continuously improving precision and recall in the IEE process, and therefore, our knowledge store. Comments from practitioners in the most recent heuristic evaluation reaffirm noticeable improvement in data coverage and representativeness. SARO should therefore ultimately serve as a comprehensive, cross-context “dictionary” and usage guide across the differences in time, location, language and culture in and beyond the EU, gradually moving us toward the ground truth in the picture of demand.

RQ3. addresses a key goal of the wider project, to identify requirements for training practitioners and entrants to the workforce, to close the gap between skill demand and existing capability. Our datastore captures demand from the viewpoint of the employment market, and therefore what decision-makers in higher level management regard as requirements for the varied roles that data scientists are expected to undertake. These decision-makers will have varying degrees of domain and/or technical expertise, or may be (technical) practitioners at senior level.

However, experienced practitioners we worked with noted that mismatch between the picture of demand and their knowledge of the field influenced their trust in the results, regardless of potential of the tools used to obtain this picture. It is necessary to correct this mismatch, to provide reliable and current results that policy- and decision-makers and educators can use to plan training. This will also ensure a guide for (practising and aspiring) data scientists in correctly identifying which skills map to their specific needs and interests and to the job market. Feeding in the viewpoint of practitioners at non-managerial level and with limited influence on decision-making is therefore critical in validating our picture of demand. An additional benefit is further information on translation and (local) usage of terminology.

The skill categorisation task requires us to rank skills, identify skill co-occurrence overall and non co-occurrence especially within a skillset. This is to ensure skills training and curricula development optimise existing knowledge and experience, and lower the risk of expending effort toward developing skills redundant for a particular role type, domain and/or industry sector. This should aid also resource allocation for training at all levels.

As an example we look again at Fig. 8, where we use parallel coordinates to examine detail in individual postings. An unexpected result was the recognition of patterns not immediately obvious in the other views – in addition to (varying) co-occurrence, skills that never co-occur. Such cases were highlighted here because the technique supports functionality for visual querying along multiple dimensions. We found, for instance, when browsing by date, that selected skills mutually excluded others. This prompted further inspection in the other views where more specific queries confirmed these results across the datasets discussed here and all others examined to that point (additional, annotated snapshots available online⁷). Inspecting the two sub-views that zoom into the *visualisation* skillset in Figs. 8(b) and 8(c), we can see that *data stories* and *d3js* never co-occur. Going back to the skill network,

¹⁶ <https://www.python.org>.

¹⁷ <http://www.tableau.com>.

¹⁸ <https://www.r-project.org>.

¹⁹ <http://nosql-database.org>.

²⁰ <http://hadoop.apache.org>.

²¹ <http://d3js.org>.

the snapshots in Figs. 4(a) and 4(b) show, respectively, the results obtained by restricting the depth of the relationships displayed in the graph to show only nodes with at least one parent in common with the focus, before expanding to the next level to reveal more distant relationships. We found this pattern of non co-occurrence was repeated.

In the significantly larger and richer datasets we find higher mention of *d3* but still very low counts of *data stories*. In a very small number of aggregates co-occurrence is recorded. Further analysis of such cases will confirm whether these are rare cases or only appear to be so due to aggregation over periods of a week or longer. We must consider also synonym search, to decrease the occurrence of false negatives. Importantly, as this task progresses this more detailed knowledge store will feed into increasingly accurate and comprehensive categorisation, and the development of training based on a validated picture of demand.

We conclude with RQ1: what is the demand landscape for data scientists across the EU?

We have highlighted frequently mentioned skills and variation in frequency across time and location; skill co-occurrence and not; and other recurring and transient patterns and trends in the data. From this a picture of the demand landscape is starting to emerge. We see a good degree of consistency across the region in terms of relative demand for the more frequently mentioned skills. Some unexpected and significant patterns are seen: in trying to determine potential movement for jobs in *statistics*, for instance, we use the (visual) country filter in the parallel coordinates plot. We see that all countries with good coverage, with the exception of France, show a peak at *statistics* along with peaks for one or more *programming skills*. (The range of filter results is available online.⁷)

Determining how to match unrelated skills to a role requires a more comprehensive dictionary and extension to the wider skill list, to filter for, say, *finance*, co-occurring with *data journalism* (and/or *data stories*) and other synonyms of both terms. Further, the job seeker may wish to restrict their search to adverts for jobs in financial institutions. Future work includes capturing metadata on industry sector — this would provide another facet along which to narrow down the focus.

Finally, our iterative, multi-perspective, task-centred process allows us to track changes in patterns and results as we obtain new data. This process supports both evaluation of our design and tools and error checking of the data and analysis results. For instance, is the atypical pattern for *statistics* in France the actual case, due to terminology usage, or an error in the input? While significant changes in patterns may be seen for a large change in the base dataset, no change at all or change inconsistent with previous data alerts to potential errors in data upload and/or violation of the data structure defined by SARO during IE.

Inconsistency in data format may also lead to queries failing to retrieve all matches, with a knock-on effect on patterns seen. We must, for instance, balance the use of regex filters that better accommodate inconsistent patterns at the cost of response time, with stricter, predefined URI patterns for skills, locations and dates that result in interactive response. Future work includes the setting of strict rules at the start of the IEE process, to minimise data loss at the (interactive) analysis stage. In related work, we are preparing to carry out large-scale validation of the skill instances being fed into SARO, to reduce reliance on human verification in the increasingly large scale annotation exercise. This includes comparing random samples of automatically annotated job postings to the results obtained by human annotators, to calculate precision and recall for the semantic annotations and the success rate of our deduplication method, key to cleaning the datastore.

This will be accompanied by the next stage of usability evaluation sessions with a wider range of users. The aim is to review

redesign based on feedback to date and updates to the user-facing tools. We will also look at the impact of increased accuracy and representativeness of the demand data especially outside the UK, and the effectiveness of the techniques we are using to manage the increase in size and complexity of our datastore.

10. Conclusion

Interactive visualisation of complex data and scenarios enables powerful, intuitive recognition of patterns and trends, aiding exploratory knowledge discovery and in-depth analysis. However, an open challenge, especially where the backgrounds and technical capability of target users vary significantly, is corresponding variation in visual literacy [1,29,31]. To address this, we follow a user- and task-centred approach that uses ontologies to abstract and aggregate dynamic, heterogeneous data into structured information spaces. This provides a backbone that structures and guides visual presentation and analysis from multiple perspectives, augmenting exploratory knowledge discovery and focused, detailed analysis by end users with different, albeit overlapping, interests and goals.

To assess whether we are able to meet the information-seeking needs of our different stakeholders and help each build a reliable and usable picture of skill demand, we examine a set of research questions that continue to guide our longitudinal study. Verification of the results obtained from our working tools is supported by feedback from target end users that highlights where our approach adds value over traditional data analysis methods. With a more complete understanding of user requirements we continue to explore additional, alternative, visual analysis techniques to satisfy these, including matrix plots in small multiples [e.g., 25] and additional time series analysis [e.g., in 4].

Employment and skill demand analysis is pertinent to the modern, data- and technology-dependent world, where skills and capabilities in a variety of industry sectors must be updated to cope with this new, invaluable source of knowledge. However, research on the topic is fairly limited. We ground our iterative process of requirements review, design, development and user evaluation with relevant research in information extraction, ontology development and visual discovery and analytics. We will make our data and results available openly where no restrictions exist due to third party input. This is to develop, also, verifiable benchmarks for assessing skill demand and existing capability, reusable in research, education and for policy-making at state level and in industry, and for building personal career pathways. We conclude with verifiable results that feed into an increasingly richer cycle of structured, shared knowledge. We continue to feed this into wider application, to the general employment and skill demand scenario, as we build a more representative and complete, data-driven picture of the skill demand landscape.

Acknowledgement

The work reported in this paper was funded by the EU project EDSA (EC no. 643937).

References

- [1] C. Ware, *Visual Thinking: for Design*, Morgan Kaufmann Publishers Inc., Burlington, USA, 2010.
- [2] J.J. van Wijk, The value of visualization, in: IEEE VIS 05: Proc., Visualization 2005, 2005, pp. 79–86, <http://dx.doi.org/10.1109/VISUAL.2005.1532781>.
- [3] A.-S. Dadzie, M. Rowe, D. Petrelli, Hide the stack: Toward usable linked data, in: ESWC 2011: The Semantic Web: Research and Applications, Proc., 8th Extended Semantic Web Conference (Part), 2011, pp. 93–107 http://dx.doi.org/10.1007/978-3-642-21034-1_7.
- [4] K. Einsfeld, A. Ebert, A. Kerren, M. Deller, Knowledge generation through human-centered information visualization, *Inform. Visualiz.* 8 (2009) 180–196. <http://dx.doi.org/10.1057/ivs.2009.15>.

- [5] R. Hervás, J. Bravo, Towards the ubiquitous visualization: Adaptive user-interfaces based on the Semantic Web, *Interact. Comput.* 23 (1) (2011) 40–56. <http://dx.doi.org/10.1016/j.intcom.2010.08.002>.
- [6] T. Costanza, European Commission report highlights ICT skills gap in Ireland, EU, 2013, available at: <https://www.siliconrepublic.com/portfolio/2013/06/13/european-commission-report-highlights-ict-skills-gap-in-ireland-eu>.
- [7] D. Donoho, 50 years of Data Science, 2015, available at: <http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf>.
- [8] D. Macrae, Why bridging the data science skills gap requires a team effort available at: <http://www.techweekeurope.co.uk/e-management/skills/bridging-data-science-skills-gap-requires-team-effort-160818>.
- [9] A.-S. Dadzie, J. Domingue, Visual exploration of formal requirements for data science demand analysis, in: VOILA! 2015: Proc., International Workshop on Visualizations and User Interfaces for Ontologies and Linked Data, 2015, pp. 1–12.
- [10] T.H. Davenport, D.J. Patil, Data scientist: the sexiest job of the 21st century, *Harv. Bus. Rev.* 90 (10) (2012) 70–76.
- [11] V. Granville, *Developing Analytic Talent*, John Wiley & Sons Inc., 2014.
- [12] S. Liu, D. Maljovec, B. Wang, P.T. Bremer, V. Pascucci, Visualizing high-dimensional data: Advances in the past decade, *IEEE Trans. Vis. Comput. Graphics* 23 (3) (2017) 1249–1268. <http://dx.doi.org/10.1109/TVCG.2016.2640960>.
- [13] G. Strawn, Data scientist, *IT Professional* 18 (3) (2016) 55–57. <http://dx.doi.org/10.1109/MITP.2016.41>.
- [14] R.W. Gehl, Sharing, knowledge management and big data: A partial genealogy of the data scientist, *Eur. J. Cultur. Stud.* 18 (4–5) (2015) 413–428. <http://dx.doi.org/10.1177/1367549415577385>.
- [15] M. Zhao, F. Javed, F. Jacob, M. McNair, SKILL: A system for skill identification and normalization, in: *Innovative Applications of Artificial Intelligence*, 2015, pp. 4012–4017.
- [16] I.A. Wowczko, Skills and vacancy analysis with data mining techniques, *Informatics* 2 (4) (2015) 31. <http://dx.doi.org/10.3390/informatics2040031>.
- [17] R. Harper, The collection and analysis of job advertisements: A review of research methodology, *Lib. Inform. Res.* 36 (112) (2012) 29–54.
- [18] M. Khobreh, F. Ansari, M. Fathi, R. Vas, S.T. Mol, H.A. Berkens, K. Varga, An ontology-based approach for the semantic representation of job knowledge, *IEEE Trans. Emerg. Top. Comput.* 4 (3) (2016) 462–473. <http://dx.doi.org/10.1109/TETC.2015.2449662>.
- [19] C. Terblanche, P. Wongthongtham, Ontology-based employer demand management, *Softw. - Pract. Exp.* 46 (4) (2016) 469–492. <http://dx.doi.org/10.1002/spe.2319>. SPE-14-0130.R2.
- [20] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damjanovic, T. Heitz, M.A. Greenwood, H. Saggion, J. Petrak, Y. Li, W. Peters, et al., *Text Processing with GATE*, Gateway Press, CA, 2011.
- [21] Occupational Information Network Resource Center, O*NET Resource Center, US Department of Labor, Employment & Training Administration Occupational Information Network Resource Center, 2015, available at: <https://www.onetcenter.org>.
- [22] J. De Smedt, M. Le Vrang, A. Papantoniou, ESCO: Towards a semantic web for the European labor market, in: LDOW 2015: Proc., Linked Data on the Web Workshop at WWW'2015, 2015.
- [23] User guide for the application of the European e-Competence Framework 3.0., European Commission, 2014, available at: http://ecompetences.eu/wp-content/uploads/2014/02/User-guide-for-the-application-of-the-e-CF-3.0CEN_CWA_16234-2_2014.pdf.
- [24] J. Bimrose, R. Wilson, S.-A. Barnes, D. Owen, Y. Li, A. Green, L. Bosworth, P. Millar, A. Holden, G. Attwell, P. Rustemeier, R. Elferink, J. Higginbottom, LMI for All Developing a Careers LMI Database: Final Report, Warwick Institute for Employment Research, Pontydyssu, Raycom, Rewired State, 2015, available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/446332/LMI_for_All_Final_report_v020715_FW.pdf.
- [25] J. Heer, B. Shneiderman, Interactive dynamics for visual analysis, *Commun. ACM* 55 (4) (2012) 45–54. <http://dx.doi.org/10.1145/2133806.2133821>.
- [26] A. Kerren, F. Schreiber, Toward the role of interaction in visual analytics, in: WSC: Proc., 2012 Winter Simulation Conference, 2012, pp. 1–13. <http://dx.doi.org/10.1109/WSC.2012.6465208>.
- [27] C. Turkay, E. Kaya, S. Balcisoy, H. Hauser, Designing progressive and interactive analytics processes for high-dimensional data analysis, *IEEE Trans. Vis. Comput. Graphics* 23 (1) (2017) 131–140. <http://dx.doi.org/10.1109/TVCG.2016.2598470>.
- [28] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, G. Melançon, Visual analytics: definition, process, and challenges, in: A. Kerren, J.T. Stasko, J.-D. Fekete, C. North (Eds.), *Information Visualization: Human-Centered Issues and Perspectives*, Springer, 2008, pp. 154–175. http://dx.doi.org/10.1007/978-3-540-70956-5_7.
- [29] D. Sacha, A. Stoffel, F. Stoffel, B.C. Kwon, G. Ellis, D. Keim, Knowledge generation model for visual analytics, *IEEE Trans. Vis. Comput. Graphics* 20 (12) (2014) 1604–1613. <http://dx.doi.org/10.1109/TVCG.2014.2346481>.
- [30] A. Inselberg, *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2009.
- [31] O. Kulyk, R. Kosara, J. Urquiza, I. Wassink, Human-centered aspects, in: A. Kerren, A. Ebert, J. Meyer (Eds.), *Human-Centered Visualization Environments*, in: *Lecture Notes in Computer Science*, 4417, Springer, Berlin Heidelberg, 2007, pp. 13–75. http://dx.doi.org/10.1007/978-3-540-71949-6_2.
- [32] T. Munzner, A nested model for visualization design and validation, *IEEE Trans. Vis. Comput. Graphics* 15 (6) (2009) 921–928. <http://dx.doi.org/10.1109/TVCG.2009.111>.
- [33] G. Elia, A. Margherita, Next-generation human resource management: A system for measuring and visualising professional competencies, *Int. J. Human Res. Develop. Manage.* 15 (1) (2015) 1–15. <http://dx.doi.org/10.1504/IJHRDM.2015.069975>.
- [34] A. Khalili, A. Loizou, F. Harmelen, Adaptive linked data-driven web components: building flexible and reusable semantic web interfaces, in: *ESWC 2016: 13th International Conference on The Semantic Web. Latest Advances and New Domains*, 2016, pp. 677–692. http://dx.doi.org/10.1007/978-3-319-34129-3_41.
- [35] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorf, H. Parkinson, The NHGRI GWAS catalog, a curated resource of SNP-trait associations, *Nucleic Acids Res.* 42 (D1) (2014) D1001–D1006. <http://dx.doi.org/10.1093/nar/gkt1229>.
- [36] D. Tarrant, S. Bullmore, M. Costello, Deliverable D1.1: Study design document, European Data Science Academy (EDSA), EC. project 643937, 2015, available at: <http://edsa-project.eu/edsa-data/uploads/2015/02/EDSA-2015-D11-Final-v1.1-forwebsite.pdf>.
- [37] E. Pietriga, C. Bizer, D.R. Karger, R. Lee, Fresnel: A browser-independent presentation vocabulary for RDF, in: *ISWC2006: The Semantic Web, Proc.*, 5th International Semantic Web Conference, 2006, pp. 158–171. http://dx.doi.org/10.1007/11926078_12.
- [38] C. Tominski, S. Gladisch, U. Kister, R. Dachselt, H. Schumann, Interactive lenses for visualization: An extended survey, *Comput. Graph. Forum* (2016) n/a–n/a. <http://dx.doi.org/10.1111/cgf.12871>.
- [39] Learning opportunities and qualifications in Europe. Information about courses, work-based learning and qualifications, European Commission, European Qualifications Framework, 2015, available at: https://ec.europa.eu/ploteus/search/site?f%5B0%5D=im_field_entity_type%3A97.
- [40] J. Volz, C. Bizer, M. Gaedke, G. Kobilarov, Silk – a link discovery framework for the web of data, in: *LDOW 2009: Proc.*, Linked Data on the Web Workshop at WWW'2009, 2009.
- [41] C. Bizer, T. Heath, T. Berners-Lee, Linked data: Principles and state of the art, in: *WWW'08: Proc.*, 17th International Conference on World Wide Web, 2008, pp. 1–40.
- [42] E.M. Sibarani, S. Scerri, C. Morales, S. Auer, D. Collarana, Ontology-guided job market demand analysis: a cross-sectional study for the data science field, in: *SEMANTICS 2017: 13th International Conference on Semantic Systems*, ACM, 2017. <http://dx.doi.org/10.1145/3132218.3132228>.
- [43] R. Isele, C. Bizer, Active learning of expressive linkage rules using genetic programming, in: *Web Semantics: Science, Services and Agents on the World Wide Web*, p. 23.
- [44] J. Nielsen, The usability engineering life cycle, *Computer* 25 (3) (1992) 12–22. <http://dx.doi.org/10.1109/2.121503>.
- [45] J. Nielsen, Usability inspection methods, in: *CHI '94: Proc.*, SIGCHI conference on Human Factors in Computing Systems, 1994, pp. 413–414. <http://dx.doi.org/10.1145/259963.260531>.
- [46] G. Ellis, A. Dix, An explorative analysis of user evaluation studies in information visualisation, in: *BELIV '06: Proc.*, 2006 AVI Workshop on BEyond time and errors: novel evaluation methods for Information Visualization, 2006, pp. 1–7. <http://dx.doi.org/10.1145/1168149.1168152>.
- [47] J. Brooke, SUS: A Retrospective, *J. Usab. Stud.* 8 (2) (2013) 29–40.