Autonomic Curation of Crowdsourced Knowledge: The Case of Career Data Management

Alina Patelli *, Peter R. Lewis *, Hai Wang *, Ian Nabney *, David Bennett †, Ralph Lucas ‡ and Alex Cole ‡
*Aston University, Birmingham, UK, Email: {a.patelli2, p.lewis, h.wang10, i.t.nabney}@aston.ac.uk
†Codevate, Birmingham, UK, Email: david@codevate.com
‡The Good Careers Guide, Email: {ralph.lucas, alex.cole}@goodcareersguide.co.uk

***Abstract*—Automatically curating knowledge that is available online is a pressing necessity, given the exponential increase in the volume of data published over the web. However, the solutions presently available are yet to reach the same level of support quality provided by human curators. This is mainly due to the fact that digital database managers do not take the expertise of the interested community into account nor exploit the underlying connections between knowledge pieces when processing user queries. We propose an approach to bridge the gap between automated curation and the one provided by human experts and implement it in the field of career data management. The resulting platform, Aviator, is based on an ontology powered autonomic manager which produces complete, intuitive and relevant answers to career related queries, in a time effective manner. We provide numeric and use case based evidence to support these research claims.**

I. Introduction

Decision making in the digital world is supported by effective knowledge processing. Given the size of the available digital data repositories, manual curation is fast becoming unfeasible. Automated query answering platforms (leveraging data from museum records [1], computerised tools for symptom based medical diagnosis inference [2], archaeological database processing [3], etc.) represent an attractive solution, however, several important issues remain unaddressed:

- The connections between different data entries are rarely and insufficiently exploited, therefore the results presented in answer to user queries lack insight and are often incomplete.
- The format that query results are presented in (commonly, lists of entries that syntactically match the search keywords) is counter-intuitive and unable to provide a coherent outlook of the relevant sub-field of the knowledge base.
- The provided results are rarely filtered based on the user's profile and interests.
- The end user has to address the problems above "manually" by explicitly searching for additional results (maybe by employing several query answering tools and collating their respective output), researching data connections and matching them against personal interests, etc. - all time consuming operations requiring intense effort.

We analyse these open problems in the career knowledge management domain, where the available data is abundant, heterogeneous, decentralised and dynamic. Yet, the workforce is expected to effectively analyse it in order to make informed decisions about the most suitable career path. For this reason, we believe the career domain offers a representative case study for investigating the proposed research question, namely how to design an automated knowledge curation platform capable of addressing all previously identified issues.

The proposed solution is Aviator, a career knowledge management system available on the GCG (Good Careers Guide) platform that stores, maintains and exposes the connections between career fields, displays query results in the form of an intuitively rendered graph (as opposed to a list), compares available knowledge against expressed user preferences and performs all these tasks automatically, thus saving a significant amount of the users' time.

The following section presents the motivation for this research and better describes the problem that we address. Section III focuses on the career knowledge domain as a representative instance of the autonomic curation context. After a brief description of Aviator's hybrid architecture (IV), the paper analyses the way that the proposed platform implements the autonomic metaphor (V). Evidence to support all research claims is provided in section VI, whereas the final sections present an analysis of related work and the paper's conclusions.

II. Motivation

Great strides have been made in recent decades to digitise information [4], [5], [6], [7], [8], as paper-based systems have been replaced by databases available over the web. In legacy paper-based systems, the role of the curator[1] was key. For example, when presented with a university student's query about "modern art", most librarians would be able to provide all the books on the official reading list. However, experienced librarians would also recommend less known yet relevant resources (websites, articles, critics' reviews) found useful by other library members on a similar academic quest. It is usually the insight provided by this sort of material that turns a good university essay into an excellent one. To provide an example from a safety-critical domain, let us think of medical staff as curators of knowledge. Decisions about patient treatment are based on the physician's core specialist knowledge

[1]A content specialist charged with an institution's collections and involved with the interpretation of heritage material - retrieved from https://en.wikipedia.org/wiki/Curator

978-1-5090-3536-6/16 $31.00 © 2016 IEEE
DOI 10.1109/ICCAC.2016.20

about human anatomy as well as on specific case studies, recent research and other clinicians' experience in similar or more loosely related domains. It is often the connections between all those sources of knowledge that enable medical professionals to formulate an accurate diagnosis.

Given the ever increasing volume of information across all fields, the pool of resources the human curator should have expert knowledge of has become intractable. The IT community's solution to this issue was to transfer all available data from a paper support to a digital one. Ideally, the entirety of the human curator's knowledge should be captured by a (library, medical, etc.) database, whereas the curation role itself would be taken over by the database manager. Realistically, that aim was achieved only to a certain extent: while the core data (library cards, patient charts, known symptoms of medical conditions, etc.) was successfully ported from hard copy versions to databases, the *experience* of human curators, namely the *connections* they were able to make between different types of knowledge, was lost along with the sense of (library, medical, etc.) *community* that used to factor into the curator's decision making process. As a result, running a query for "modern art" in a digital database will no longer return the additional resources that do not exactly match the search keyword but that the human librarian had knowledge of. Similarly, a diagnosis based only on the results returned by a medical symptoms' database will not account for specific yet relevant cases that human doctors would know of and be able to interpret.

One way to address this became available with the dawn of Web 2.0 [9], a reinvention of the classic World Wide Web, where online content is curated by non-expert users. This is done by annotating web resources with *tags*, usually as simple as words, that concisely capture one aspect of the online content. For instance, a digital print of a Monet painting could be tagged with "water lily" to describe its theme and "blue" to refer to the predominant colour. This approach to online content management has proven very attractive, with websites such as youtube, Delicious, Flickr [10] and Pinterest [11] gaining increased popularity. The immediate advantage is that separate resources are connected via user tags, thus reinstating a sense of *community*, on the one hand, as well as allowing for better, more powerful search algorithms, on the other hand (if those additional library resources existed on a Web 2.0 website and were tagged with "modern art", they would be included in the query results alongside the traditional matches).

However, an important problem remains. Left unchecked, that is, under the exclusive control of the individual user, tags may quickly become ambiguous (a different user may tag the same Monet painting with "pond flowers"), conflicting (a given viewer may be of the opinion that the predominant colour in the print is "green", not "blue") or incorrect (the print may be wrongfully tagged with "Manet" instead of "Monet"). Thus, the added value brought by community curation turns against itself and sabotages the powerful search algorithms it was meant to support. One possible solution lies with the Semantic Web [12], another iteration in the World Wide Web's transformation, where user tags are regulated by an ontology [13], [14], [15]. An ontology stores concepts and the properties connecting them in the form of a graph, expressed in the light logic formalism of RDF (Resource Description Framework) [16]. The lead advantage provided by an ontology in the online curation context is disambiguation: every concept is represented alongside its synonyms (maintained by an expert, by the larger community, or, ideally, by both) that can be used by search algorithms to identify equivalent tags and eliminate conflicting ones.

Ontologies offer intrinsic support for some of the challenges identified in the introduction (they store synonyms, therefore are capable of running "richer" queries, with a better yield and they are structured as graphs - with concepts in the role of nodes and properties acting as edges - knowledge models that are intuitive and easy to explore in order to get a comprehensive perspective of the relevant sub-field). However, taking user preferences into consideration in order to produce relevant query results implies some supplementary logic that ontologies do not provide native support for. Also, running complete queries, allowing graph exploration and filtering results based on user profile are all tasks that need to be executed automatically, which is again beyond the core capabilities of ontologies as standalone platforms.

III. The Career Management Scenario

In this paper, we tackle the above challenges in the domain of career management platforms, which are a typical example of a knowledge base in transition from paper to the digital world. Further, effective career management platforms are crucial tools in providing support for informed decision making for the entire workforce.

In its current form, the online career space comprises knowledge from three sources:

- **experts** (National Careers Service, relevant Wikipedia pages, etc.) providing general information about professional fields and the way they connect to each other, for instance, the fact that "chemistry" is a sub-field of "science"
- **providers** of either education (universities publishing academic requirements for pursuing a given career, HESA[2] maintaining the latest JACS[3] list) or jobs (company websites offering specific career / role description, job adverts published via third party websites, such as indeed.co.uk)
- **explorers** of online, career relevant content in search of a new job or a better understanding of their professional prospects and assigning tags or writing reviews in the process.

The only form of automatic career knowledge management available for explorers is provided by job search engines

[2]Higher Education Statistics Agency - https://www.hesa.ac.uk/

[3]Joint Academic Coding System - https://www.hesa.ac.uk/component/content/article?id=1787

(e.g., indeed.co.uk, jobs.ac.uk), as illustrated in Fig. 1. These take in one or more keywords and produce a list of job adverts based on syntactically matching the provided keywords against the text description of the jobs. Besides the semantic incompleteness of the results (relevant jobs may be omitted from the list if published under a synonym of the search keyword that the explorer is unaware of), such search engines disregard the first and third knowledge sources altogether. The *connections* between career concepts as well as the explorer *community* output (in the form of tags and reviews) are thus obscured. Consequently, explorers take the curator's role upon themselves and sift through HESA content to match their academic credentials against job requirements, read generalist web pages with broad scope information about each role in the result list and consult other explorers' reviews and comments in order to make an informed decision about applying for a given job or not.

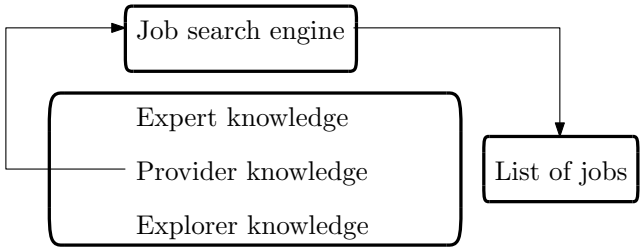

Fig. 1. The career space managed by job search engines - expert and explorer knowledge is ignored

We introduce Aviator[4], a career management platform available on the Good Careers Guide platform that allows explorers to tag career resources with concepts from an ontology and benefit from each other's expertise. Aviator addresses the previously identified issues by:

- offering **completeness** in the sense that all synonyms of a career concept known to the system will be considered when compiling the associated list of jobs
- providing **perspective** by displaying the career concepts relevant to the user query as well as their connections (the latter are unavailable in the classic list format that job search results are displayed in)
- enhancing **relevance** via collecting all the tags that a registered user annotated online resources with and using them to generate a personal ontology - this can be compared against the ontology of the ideal candidate for a given role, thus allowing jobs that do not match the user's career profile to be filtered out
- saving **time** gained by having queries answered in a complete and automatic fashion, rather than manually curating the relevant knowledge.

At an architectural level, in Aviator, the role of the curator is fulfilled by an autonomic manager (Fig. 2), where the

[4]https://gcg-test.codevate.com - log in with user name "johnsmith" password "gcgtesting"

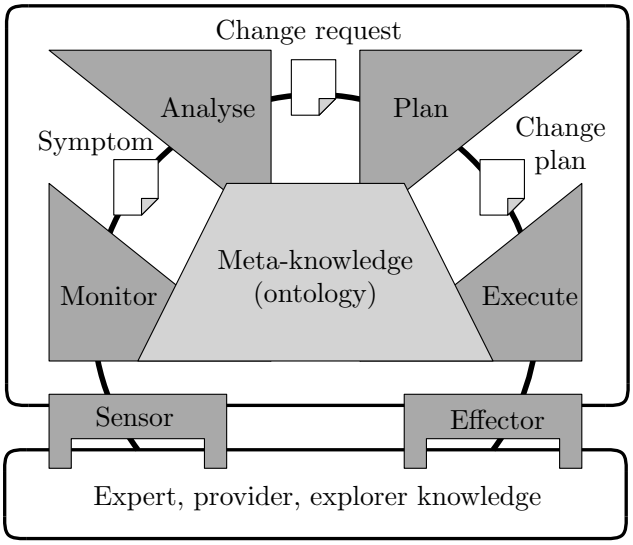

Fig. 2. The career space managed by an autonomic manager - adapted from [17]

knowledge component of the MAPE-K (monitor analyse plan execute - knowledge) loop is an ontology. The *monitor* collects information from **providers** (new jobs posted on indeed.co.uk) and **explorers** (tags, reviews, ontology edits). The *analyse* module verifies the consistency of the underlying knowledge base, accepting/rejecting edits accordingly, and maintains the list of tags used by every registered system user. The *plan* component runs either a simple query (to retrieve the segment of career knowledge that the user is interested in exploring) or a compound one (essentially, a separate query for each tag) to compile a personal ontology. The results of the query are displayed in the *execute* phase as a graph in the system's visualiser. The *knowledge* informing the operation of the MAPE loop is represented in the form of an ontology, initially extracted from a legacy document containing **expert** knowledge about career concepts, their properties (synonyms and connections) and the relevant JACS codes. The ontology is maintained via user edits and displayed (in segments) in response to user queries.

IV. Architecture Description

The two main Aviator components (illustrated in Fig. 3) are the *web server* hosting the user interface and the *ontology server* providing a feature rich semantic platform capable of running user requested services (e.g., incremental graph exploration, graph editing, personal ontology generation).

The *searcher* takes a keyword from the user and matches it against ontology concepts. The ontology nodes related (via a maximum of two links) to the matching concept are displayed in the *visualiser* (the visualisation plug-in used by Aviator is Cytoscape[5]). To illustrate the process, Fig. 4 shows the result of the search for keywords "quantitative methods".

[5]http://www.cytoscape.org/

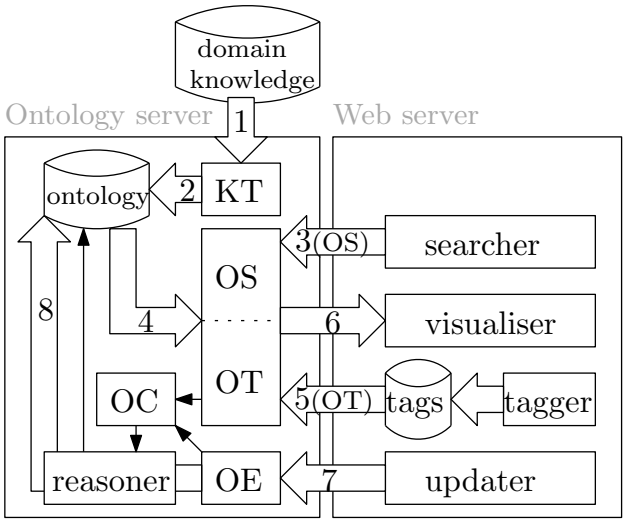

Fig. 3. Aviator architecture: line arrows represent "uses" relationships between components, block arrows illustrate the data flow through the system

The website *tagger* (Fig. 5) displays as a menu to the left of the page being annotated and allows users to assign tags (i.e., concepts from the ontology) to its content. If the page has been previously reviewed by other users, that information is available in the menu as well. Every ontology concept used as a tag by a specific user is stored in the *tags* collection.

The *updater* (available only to administrators) allows the editing of the ontology by adding/deleting concepts and their connections through a dedicated graphical interface. Fig. 6 illustrates the process of creating a new link (asserting a new ontology property) connecting "biological computing" (the *Source*) to "applied biological sciences" (the *Destination*). The system displays a list of existing ontology concepts currently related to the source and destination (e.g., "biological computing" currently has two parents, "biology" and "applied computing"), which is meant to inform the user's decision with respect to the most appropriate type of link to assert. In the example in Fig. 6, "biological computing" is made the child of "applied biological sciences" (the other two options are parent or sibling).

The *ontology* is expressed in OWL[6] and extracted from a data repository (marked *domain knowledge* in Fig. 3) provided by a domain expert. Besides career related concepts, the ontology also stores relevant subject identifiers in the Joint Academic Coding System (JACS)[7], thus making the Aviator ontology compatible with HESA and UCAS standards for UK higher education. The relationships between career nodes are expressed via three semantic properties, namely `hasParent`, `hasSibling` and `hasSynonym`. The ontology is created by the knowledge translator from expert provided career data (flow arrow 2 in Fig. 3) and is modified by the ontology editor, which is in charge of implementing user updates (flow arrow 8 in Fig. 3). Ontology content is fed into the ontology segmenter and the ontology tailor, responsible with creating custom graph views in response to user requests (flow arrow 4).

The *knowledge translator* (*KT* in Fig. 3) populates the ontology by transforming domain knowledge into semantic contents. This process corresponds to data flow stages 1 and 2 in Fig. 3.

The *ontology segmenter* (*OS*) and the *ontology tailor* (*OT*) are represented in the same block in Fig. 3, as they share input 4 (feeding from the ontology) and output 6 (displaying the generated results in the visualiser). Input 3, namely the keyword used in the search, is specific to the OS only. Input 5, that is, the collection of tags (ontology concepts) that the current user annotated career webpages with, feeds exclusively into OT. In terms of actual operation, OS matches a search keyword against an existing ontology concept *c* (input 3) and runs a DL query over the ontology (input 4) to extract a set of nodes related (over a maximum of two connections) to *c* via `hasParent` or `hasSibling` properties. The resulting ontology segment is fed into the visualiser (output 6) where it is displayed as a graph. The *ontology tailor* (*OT*) performs the same operation as the OS, only in batch mode, once for every element in *tags* (input 5). Each query will produce a graph, their ensemble forming the current user's personal ontology (output 6). These are useful for job seekers as they represent visual descriptions of their professional interests, in other words, a history of their job related web browsing.

The *ontology editor* (*OE*) receives the modification suggested by the user (e.g., a concept/link addition/deletion) through the updater (input 7) and asserts it in the ontology (output 8). There are three types of edits currently available through the ontology administration interface: turning a node into a synonym and vice versa, adding a new concept (the parents, siblings, children and synonyms of the added node need to be specified as well) and adding/deleting a link. The addition of a new link via the updater (making "biological computing" a child of "applied biological sciences") is illustrated in Fig. 6. Given the sensitive nature of the edit operation (that allows end users to modify the knowledge base), the updater is currently only available to administrators.

The *reasoner* is meant to maintain the logical consistency of career related knowledge as well as infer new knowledge to support the ontology search process. The *ontology classifier* (*OC*) is the component in charge of deploying the reasoner whenever necessary (e.g., before committing changes to the ontology, to ensure logical consistency is maintained).

V. Autonomic Curation

This section explains how autonomic curation of crowd-sourced knowledge is implemented in the context of the above architecture.

While the system is running, the components in Fig. 3 interact in a way that can be described as a MAPE-K loop (Fig. 2). Thus, the Aviator platform may be viewed as an autonomic system [18], where the careers' knowledge space (authored by experts, providers and explorers) is the managed resource

[6]http://www.w3.org/2001/sw/wiki/OWL

[7]https://www.hesa.ac.uk/jacs3

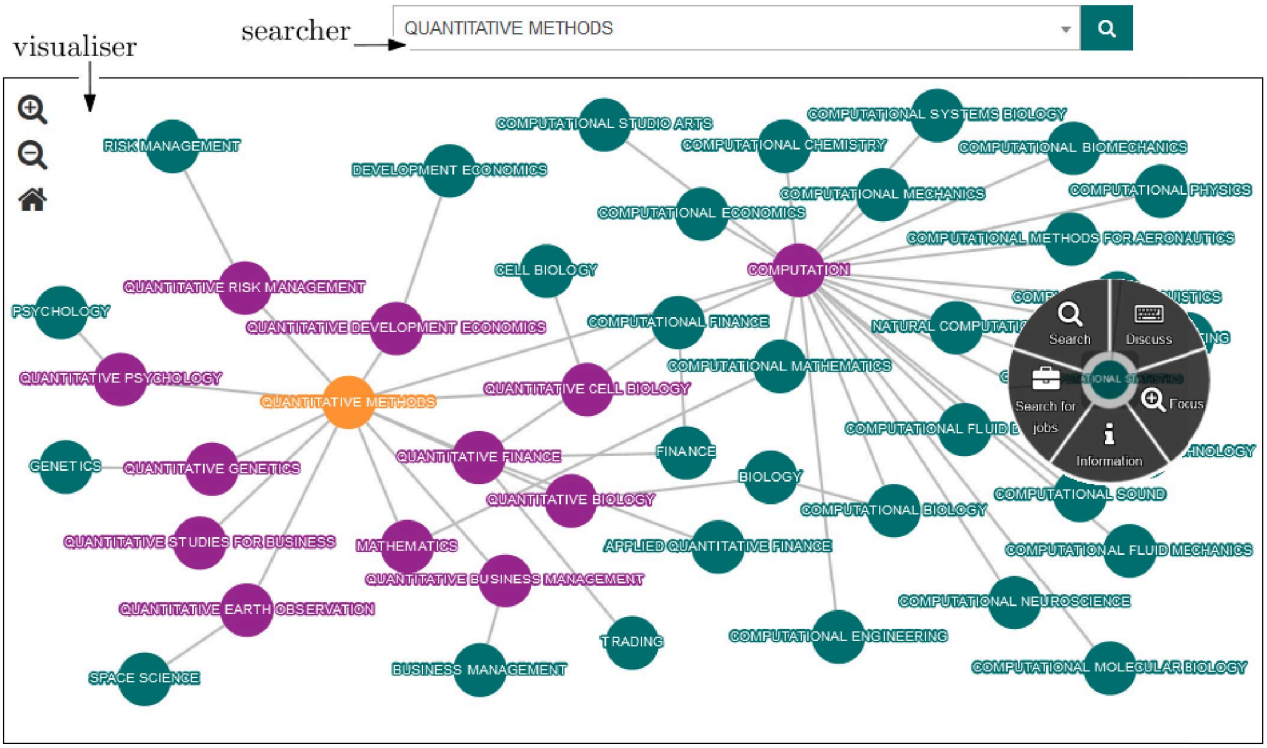

Fig. 4. Aviator: *searcher* and *visualiser* components

(curated knowledge) and the remaining components of the ontology server make up the autonomic manager (curator). The goal of the system is to offer completeness of search results, provide a perspective of the wider careers field, enhance the relevance (with respect to personal professional interests) of career related searches and save the user a significant amount of time by automating all above tasks (see section III). These four sub-goals represent a mix of qualitative and quantitative benefits and it makes little sense to aggregate them in a conventional objective function. However, an initial evaluation of these sub-goals is provided in section VI.

The components of the autonomic manager are described in the following.

A. Monitor

The Aviator system monitors:

- S1 provider knowledge (new jobs posted on indeed.co.uk)
- S2 explorer knowledge (new tags utilised by registered users to annotate, via the tagger in Fig. 5, online career resources)
- S3 expert knowledge (expressed by editing the ontology, namely adding, deleting or redefining ontology concepts and properties via the updater in Figure 6).

The web server provides the "software sensors" to capture changes in the three knowledge sub-spaces and pass them to the appropriate ontology server components. The monitoring behaviour of the Aviator system is described by the following pseudo-code.

```
monitor(Sensor[] sensors)

while(true)
  foreach s in sensors do
    if s.isActive() then
      analyse(s);
    end if
  end for
end while
```

List `sensors` contains S1 through S3, method `isActive()` returns whether sensor `s` has detected a change and `analyse()` is the method that represents the analyse phase.

B. Analyse

Analysis mostly consists in discriminating between the several types of monitored requests (via the `getType()` method), translating the sensor data (retrieved by `getOutput()`) to the right format and selecting the appropriate plan. The second `analyse()` input represents the author of the change detected by sensor `s`. The pseudo-code describing the analysis phase is presented below.

```
analyse(Sensor s, Author a)

switch(s.getType())
  case S1:
    keywords = parse(s.getOutput());
    for each k in keywords do
      c = getConcept(onto)
      assign(s.getOutput(), c);
```

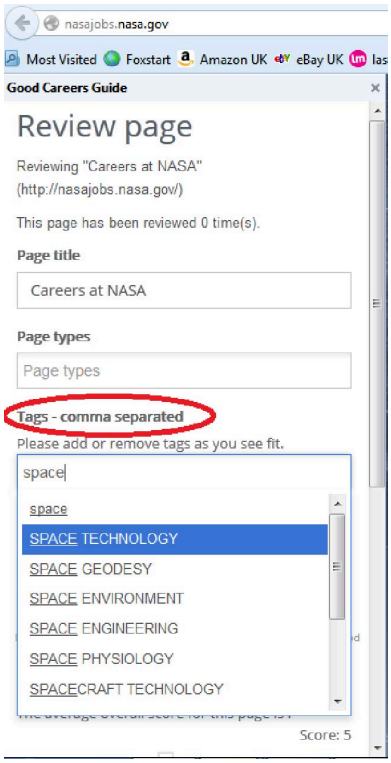

Fig. 5. Aviator: the *tagger* component used to annotate nasajobs.nasa.gov

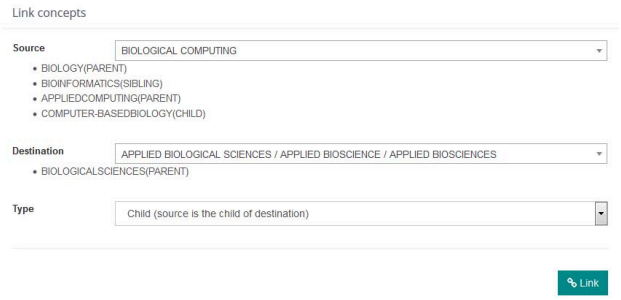

Fig. 6. Aviator: the *updater* component

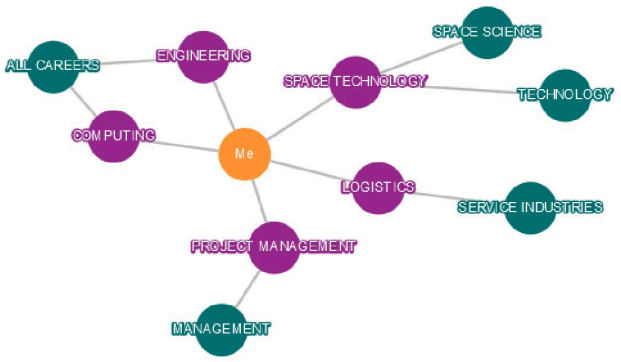

Fig. 7. GCG Aviator: a personal ontology example

```
9       end for
10    case S2:
11      clearVisualiser();
12      tags = s.getOutput();
13      plan1(tags);
14    case S3:
15      if author.isAdmin() then
16        update = s.getOutput();
17        (isConsistent, tempOnto) = plan2(update);
18        if isConsistent == true then
19          onto = tempOnto;
20        end if
21      end if
22  end switch
```

In the case of S1, the sensor data returned by method `getOutput()` is the new job post. Method `parse()` extracts the post's `keywords`, which are then matched against ontology concepts (line 7). The new job post is included in the list associated to each previously identified concept `c` (line 8). Method `clearVisualiser()` resets the graph display (Fig. 4), whilst `plan1()` and `plan2()` refer to the selected plans. The administrator privileges of the detected change's author are either confirmed or invalidated by method `isAdmin()`.

C. Plan and Execute

In order to generate a user's personal ontology (the responsibility of OT in Fig. 3), it is necessary to retrieve all concepts used as `tags` throughout the user's web exploration history and run a semantic query for each of them. The latter task is performed by the OS (see Fig. 3) by delegating to the reasoner. The query output is the matching concept's vicinity (`view` in the `plan1` pseudo-code) or null (in case the keyword did not match a concept). Method `display()` compounds the views generated for each tag and displays them in the visualiser. The same plan is used to explore the general career ontology (the centrepiece in Fig. 2), in which case the `tags` input is replaced by the keyword typed in the searcher and the `for` loop on line 3 becomes unnecessary as lines 4 and 5 need only be executed once.

```
1  plan1(Tag[] tags)
2
3  for each tag in tags do
4    view = OS(tag);
5    display(view);
6  end for
```

Edits formulated by users are performed on a temporary copy of the ontology which is afterwards submitted to the reasoner for consistency checking. The reasoner output will then be analysed (case S3 in `analyse()`) and acted upon by either committing the changes to the public ontology or dismissing them altogether. The associated plan is:

```
1  plan2(Change c) returns boolean isConsistent,
2                              Ontology tempOnto
3
4  (isConsistent, tempOnto) = runReasoner(change);
5  return (isConsistent, tempOnto);
```

Line 4 above describes the function of the OE. The reasoner will return the updated ontology along with a flag indicating whether logical consistency is met or not.

The "software effector" executes the steps of `plan2` on the ontology and those of `plan1` on the front-end display, namely the visualiser in Fig. 3 (the latter operation is supported by the Cytoscape plugin). Specifically, the effector executes one of two management actions: commits changes to the ontology after a reasoner-approved edit or displays a sub-view (either single query output or personal ontology) of the graph in the visualiser.

VI. Evaluation

The claimed benefits of autonomic curation of crowdsourced career knowledge are completeness, perspective, relevance and time. This section evaluates Aviator's capacity of practically realising these four benefits. A brief analysis of the platform's realtime operation is also provided.

A. Completeness

Let us assume that a user is interested in getting a job in *advertising*. The results list provided by indeed.co.uk for that keyword used on its own[8] contains 864 job adverts. However, the ontology features several synonyms for the concept "advertising" (Fig. 8), which, when considered together (via "Advanced search", in the textbox labeled "With at least one of these words"), form a query that yields a result list with 882 entries. To get access to these 18 extra jobs, the user would need to manually compile a list of all "advertising" synonyms, a task successfully automated by Aviator. The difference is even more striking if, by chance, the user searches for "creative director", which produces 21 results (thus, a negative difference of 861 jobs relative to the Aviator query).

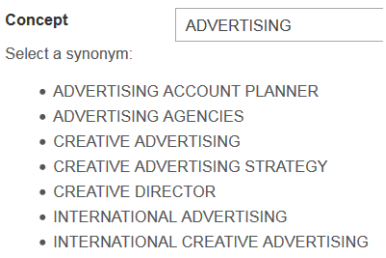

Fig. 8. Synonyms for "advertising" in the Aviator ontology

Since the lists of jobs found for each of the "advertising" synonyms are overlapping, the complete query performed by Aviator merely broadens the result set (the jobs at the intersection of the results' lists would be identified by any of the individual queries). However, the added benefit brought by Aviator queries becomes more evident in cases where there is no overlap between the lists obtained for each career synonym. For instance, "business intelligence" and "business information management", taken as two individual queries, yield 252 and, respectively, 3 results. When aggregated in a single query (in the Aviator ontology, they are synonyms), the result set contains all 255 results, showing that the lists generated for the individual queries are completely distinct. In such a situation, without Aviator support, a user who is unaware of the synonymy relationship would be deprived of the entirety of jobs associated to either one field or the other. Thus, Aviator replicates the domain expertise of a curator, in automatically referring the user to other jobs of interest, even though they were outside of the user's specific search.

[8]All indeed.co.uk search results in this paper refer to searches conducted in May 2016 in the Birmingham area with a radius of 25 miles.

B. Perspective

Entering a keyword in the Aviator searcher will produce a graph containing relevant ontology nodes *as well as their connections* (this can be tested by navigating to https://gcg-test.codevate.com/explore and exploring the graph returned for any preferred ontology concept). Displaying the output of a query as a graph provides the user with an overarching perspective of the field of interest, which is not available (or, at best, severely obscured) when presenting query results in the form of a list (such as relevant job adverts are formatted by indeed.co.uk, for instance). Moreover, from a human computer interaction perspective, research shows that displaying knowledge as graphs is more informative than lists (indented trees) [19], [20]. Another, indirect, advantage of displaying a career connectivity map rather than a list is the possibility of uncovering new, potentially relevant careers that the user may not have considered otherwise. For instance, searching for "biophysics" produces a graph featuring the expected connections ("biology" and "physics" are parents of the search topic) as well as unexpected ones ("astrobiology" is also a child of "biology" and "physics" and may be of interest, even if merely borderline, to a person with expertise in "biophysics"). This new connection is another example of something that might frequently have been pointed out by an expert curator, but would have been difficult to spot in a list of job adverts returned from a keyword search.

C. Relevance

A personal ontology comprises all Aviator concepts used as tags by a given registered system user. Those concepts, along with their one-step neighbours, are connected to a central node (labeled "Me" in Fig. 7) and displayed in the visualiser (https://gcg-test.codevate.com/explore/personal). Besides acting as a personal career profile (reflecting users' professional interests throughout their use of Aviator), a personal ontology can serve as a benchmark when searching for jobs.

Specifically, let us assume that a fictional organisation, CompX, registers an Aviator account and publishes the personal ontology for the ideal candidate they would like to hire to fill a given role. In this example, CompX's personal ontology contains all the nodes in Fig. 7 (where "Me" is also replaced with "CompX"), apart from the "project management" branch. Let us also assume that John, a young engineer and owner of the personal ontology in Fig. 7, wants to find out if CompX's job offer is a good fit for him. Upon seeing the degree of overlap between his personal ontology and the one published by CompX, John can make an informed decision with respect

TABLE I
ONTOLOGY OPERATIONS DURATION

Operation	Runtime [s]
KT(1→2)	2
KT(2→1)	3.6
OS (4)	0.9

(a) One-off operations

Node links	Query Runtime [ms]
<50	132
50 - 100	532
>100	1231

(b) Queries

to either applying for the job or not. Specifically, John may decide to keep looking for roles that also require project management skills or "sacrifice" his interest in that field and apply for the CompX job. By performing this comparison against a wide range of ideal candidate ontologies, users are supported in applying only for those jobs that are relevant to their professional interests.

D. Time

Using Aviator to manage career data saves the end user's time mainly in two ways. Firstly, by storing a list of synonyms for each ontology concept, Aviator supports broader scope queries. To achieve a similar result, the explorer would have to manually compile the synonym list from various sources, a time consuming activity thus rendered unnecessary. Secondly, comparing personal ontologies against those describing ideal candidates for available roles protects the end user from having to explicitly investigate the overlap between a given job description and personal professional interests.

E. Scalability

In support of Aviator's scalability to knowledge bases of different sizes, we provide numerical evidence of the platform's realtime performance. The main semantic operations and their execution times are listed in Table I. Sub-table Ia shows the duration of all operations that get run only once per user session: knowledge translation, both from xlsx to RDF/OWL, KT(1→2), and vice-versa, KT(2→1), and ontology loading prior to running a semantic query, OS(4). The associated execution times are measured on a DigitalOcean server: 4 CPUs @ 8GB RAM for the ontology server and 2 CPUs @ 4GB RAM for the web server. Sub-table Ib lists average query execution times measured for three types of nodes: loosely connected (with less than 50 first and second order children, parents and siblings), well connected (between 50 and 100 related concepts) - this is the category 75% of ontology nodes fall under - and highly connected (with over 100 neighbours). The number of node neighbours is the only parameter considered since it has the most significant impact of the computational cost of semantic queries. The values in column two represent the average execution times for 100 ontology nodes from each of the three categories. As expected, query runtime increases as node connectivity goes up, however, for 75% of all possible queries, the execution time is under one second. The most computationally expensive operation performed by Aviator, ontology classification, employs the FaCT++ reasoner [21] and takes, on average, 91s.

VII. RELATED WORK

Autonomic curation of online knowledge has received limited attention from the research community. Contributions usually target specific applications, such as curating meta-data associated to digital records with the purpose of cataloguing those for long-term storage [22], [23]. A similar idea to that underpinning Aviator is used to allow the community-led curation of artworks in a digital gallery [24], however, the curation process has to do with the users' artistic preference rather than semantic content. None of these contributions use an ontology to store the knowledge piece of the autonomic manager nor give any insight into the MAPE-K loop they employ.

On the other hand, the area of online career support has proven more popular. Several career support platforms make use of ontologies to store and maintain relevant knowledge. The Enterprise Ontology [25] stores the vocabulary for the business enterprise domain. Career ontologies in the ICT field are either aimed at facilitating the access of school leavers to the ICT curriculum, jobs, skills, etc. [26] or focus on improving the ontology search process (complete with a metric for measuring the relevance of ontology concepts with respect to user keywords) [27]. Other approaches [28], [29] build ontologies from secondary school student data (psychological test results and exam marks) as well as expert provided career data (only broad domains, such as literature, humanities, mathematics, are considered). By matching student data against career requirements, the systems will recommend the best fitting field of professional practice [28] or the necessary courses to take in order to meet a given degree's promotion criteria [29]. A career advice platform is used as a case study to illustrate knowledge maturing [30], namely the process of transforming highly conceptualised entities into formal, explicitly linked concepts. The platform suggests the inclusion of knowledge graph visualisation components and analyses the benefits of effective retrieval of relevant information, yet the discussion is exclusively carried out at a design level. Another approach [31] uses knowledge graphs to allow an easier understanding of mathematical concepts and mainly focuses on how to manually build the graph rather than extract it from a legacy repository.

We also analyse a body of work dealing with *knowledge graphs*, not necessarily strictly related to the careers domain, but inherently relevant to Aviator (ultimately, a knowledge graph in its own right). Wikipedia is one of the leaders in this category, given the successful exploitation of Wikidata, an ontology used to extract connections between concepts in various languages. Since the data is not explicitly exposed to the end-user, there is an overall scarcity of programmatic interfaces to the Wikipedia ontologies [32]. The Google Knowledge Graph [33] adds a semantic layer to the classic search engine (the right hand side menu next to the Google search results list is functionally similar to an Aviator sub-graph). However, clicking on a node only displays local information, without expanding the search to another view. On the other hand,

Google Knowledge Graph is a powerful, broad spectrum tool, whereas the Aviator ontology is topical in the field of careers, thus better suited to resolve specific queries. A study [34] of how knowledge diversity influences the retrieval of specific ontology data, in the presence of a size restriction, has a possible application for phase two of our platform. Link strengths may be used to define the distance between concepts, thus providing a metric to measure diversity. An excellent survey of techniques for building, mining and expanding knowledge graphs [35] is exemplified on Freebase, the ontology behind Google's Knowledge Graph. Graph Query by Example [36] is a system that uses knowledge samples as a starting point for building queries, in an effort to simplify their structural complexity (illustrated on Freebase and DBpedia). Finally, various relation extraction techniques are suggested [37] in an attempt to transform data from linguistic resources such as WordNet into knowledge graphs.

In summary, the reviewed ontology based career support platforms target narrow professional domains (such as ICT), provide guidance that mainly consists in advanced semantic search features and allow limited support for incorporating non-expert input. In contrast, Aviator employs an ontology spanning over several career fields and offers a rich set of features (career graph navigation and editing, exploration history tracking, etc.). Knowledge graph contributions provide some visualisation of the underpinning ontology, yet are either too specific [37], [35], [34] or designed for too broad a domain of applications (Google Knowledge Graph) to match the flexibility (node expansion, community edits) and customisation (personal ontologies) of Aviator.

VIII. Conclusion

Aviator is a career knowledge management platform that is relevant to the more general context of automatic knowledge curation. It exposes the underlying data in navigable, editable views of manageable size, accepts external edits after consistency verification and offers personalised snapshots of users' engagement with the system. This enables Aviator to provide a flexible (as broad or as specific as desired) *perspective* of the field, as opposed to classical career advice platforms where the subtle connections between professions, jobs, educational resources, etc. are obscured by the sheer volume of provided data. Additionally, the Aviator ontology reflects the views of a larger *community* than that of domain experts and also provides the means to *customise* the career researching experience of its end users.

Aviator is powered by a hybrid architecture where semantic tools address knowledge consistency and retrieval issues, whilst the autonomic components manage ontology changes. In this setting, the ontology has several roles: it aligns community curated information (one form of alignment is storing synonyms for each ontology concept), it supports the rendering of relevant knowledge in the form of a navigable graph and it ensures the logic correctness of the knowledge model, via reasoner performed classification.

Future work will be directed towards gathering and analysing Aviator user data (tag usage, graph exploration trends, personal ontology evolution). This will enable further platform validation as well as help study part of the career-interested community's dynamics. The latter outcome may prove useful for formulating education/training policies and labour force recruitment strategies. The second planned development is related to the introduction of numerical weights to model the strength of node connections, e.g., "science" is tightly connected to "physics" (link strength 100) but loosely connected to "astrology" (link strength 5). The analyse phase will use these weights to display only the nodes connected via strong links to a search keyword. In a broader context, the possibility of applying the autonomic knowledge curation approach embodied by Aviator to other representative domains (e.g., Pinterest) will also be investigated.

Acknowledgment

The authors would like to thank Good Careers Guide, Codevate and Capgemini for their continuous support.

References

[1] M. Taheriyan, C. Knoblock, P. Szekely, J. L. Ambite, and Y. Chen, "Leveraging linked data to infer semantic relations within structured sources," in *Proceedings of the 6th International Workshop on Consuming Linked Data (COLD 2015)*, 2015.

[2] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach," *Expert Systems with Applications*, vol. 40, no. 1, pp. 96–104, 2013.

[3] M. S. R. Simbulan, "The challenge of managing archaeological databases: Some issues and concerns," *Hukay*, vol. 19, 2014.

[4] J. Gemmell, G. Bell, and R. Lueder, "Mylifebits: a personal database for everything," *Communications of the ACM*, vol. 49, no. 1, pp. 88–95, 2006.

[5] P. M. Allison, "Dealing with legacy data-an introduction," 2008.

[6] W. Ferguson, "New geothermal data system could open up clean-energy reserves," 2013.

[7] C. Fairall, H. Edmunds, and D. Cave, "Bfi national archive: Digital workflow for the preservation of digital cinema packages," *Journal of Digital Media Management*, vol. 2, no. 2, pp. 127–136, 2013.

[8] P. Pandey and R. Misra, "Digitization of library materials in academic libraries: Issues and challenges," *Journal of Industrial and Intelligent Information Vol*, vol. 2, no. 2, 2014.

[9] T. o'Reilly, *What is web 2.0.* " O'Reilly Media, Inc.", 2009.

[10] S. J. Cunningham, "Mining flikr for museum feedback: Case study on the qatar islamic museum of art," in *Qatar Foundation Annual Research Conference*, no. 2013, 2013, pp. SSHP–035.

[11] C. Hall and M. Zarro, "Social curation on the website pinterest. com," *proceedings of the American Society for Information Science and Technology*, vol. 49, no. 1, pp. 1–9, 2012.

[12] T. Berners-Lee, J. Hendler, O. Lassila *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[13] V. Lopez, M. Fernández, E. Motta, and N. Stieler, "Poweraqua: Supporting users in querying and exploring the semantic web," *Semantic Web*, vol. 3, no. 3, pp. 249–265, 2012.

[14] D.-E. Spanos, P. Stavrou, and N. Mitrou, "Bringing relational databases into the semantic web: A survey," *Semantic Web*, vol. 3, no. 2, pp. 169–209, 2012.

[15] E. Blomqvist, "The use of semantic web technologies for decision support–a survey," *Semantic Web*, vol. 5, no. 3, pp. 177–201, 2014.

[16] R. Cyganiak, D. Wood, and M. Lanthaler, "Resource description framework (rdf): Concepts and abstract syntax," *World Wide Web Consortium, Jan*, 2013.

[17] IBM, "An architectural blueprint for autonomic computing," IBM, Tech. Rep., 2005. [Online]. Available: http://www-03.ibm.com/autonomic/pdfs/AC%20Blueprint%20White%20Paper%20V7.pdf

[18] J. Kephart and D. Chess, "The vision of autonomic computing," *Computer*, vol. 36, no. 1, pp. 41–50, 2003.
[19] B. Fu, N. F. Noy, and M.-A. Storey, "Indented tree or graph? a usability study of ontology visualization techniques in the context of class mapping evaluation," in *International Semantic Web Conference*. Springer, 2013, pp. 117–134.
[20] V. Kasyanov and E. Kasyanova, "Information visualisation based on graph models," *Enterprise Information Systems*, vol. 7, no. 2, pp. 187–197, 2013.
[21] D. Tsarkov and I. Horrocks, "Fact++ description logic reasoner: System description," in *International Joint Conference on Automated Reasoning*. Springer, 2006, pp. 292–297.
[22] I. Subotic, L. Rosenthaler, and H. Schuldt, "A distributed archival network for process-oriented autonomic long-term digital preservation," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2013, pp. 29–38.
[23] W. Allasia, F. Falchi, F. Gallo, and C. Meghini, "Autonomic preservation of access copies of digital contents,," *Proceedings of The Memory of the World in the Digital Age: Digitalization and Preservation*, 2012.
[24] K. Hazelden, M. Yee-King, M. d'Inverno, R. Confalonieri, D. De Jonge, L. Amgoud, N. Osman, H. Prade, C. Sierra *et al.*, "Wecurate: Designing for synchronised browsing and social negotiation," in *The first International Conference on Agreement Technologies*, 2012.
[25] M. Uschold, M. King, S. Moralee, and Y. Zorgios, "The enterprise ontology," *The knowledge engineering review*, vol. 13, no. 01, pp. 31–89, 1998.
[26] K. L. Chin and E. Chang, "A sustainable ict education ontology," in *Digital Ecosystems and Technologies Conference*, 2011, pp. 350–354.
[27] P. Singto and A. Mingkhwan, "Semantic searching it careers concepts based on ontology," *Journal of Advanced Management Science*, vol. 1, no. 1, 2013.
[28] M. A. Alimam, H. Seghiouer, and Y. Elyusufi, "Building profiles based on ontology for career recommendation in e-learning context," in *Multimedia Computing and Systems (ICMCS), 2014 International Conference on*. IEEE, 2014, pp. 558–562.
[29] C.-Y. Huang, R.-C. Chen, and L.-S. Chen, "Course-recommendation system based on ontology," in *Machine Learning and Cybernetics*, vol. 3. IEEE, 2013, pp. 1168–1173.
[30] N. Weber, K. Schoefegger, J. Bimrose, T. Ley, S. Lindstaedt, A. Brown, and S.-A. Barnes, "Knowledge maturing in the semantic mediawiki: A design study in career guidance," in *Learning in the Synergy of Multiple Disciplines*, ser. Lecture Notes in Computer Science, U. Cress, V. Dimitrova, and M. Specht, Eds. Springer Berlin Heidelberg, 2009, vol. 5794, pp. 700–705.
[31] B. Zwaneveld, "Structuring mathematical knowledge and skills by means of knowledge graphs," *International Journal of Mathematical Education in Science and Technology*, vol. 31, no. 3, pp. 393–414, 2000.
[32] C. M. Torsten Zesch and I. Gurevych, "Extracting lexical semantic knowledge from wikipedia and wiktionary," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, B. M. Nicoletta Calzolari, Khalid Choukri, Ed., 2008.
[33] T. Steiner, R. Verborgh, R. Troncy, J. Gabarro, and R. Van de Walle, "Adding realtime coverage to the google knowledge graph," in *11th International Semantic Web Conference (ISWC 2012)*, 2012.
[34] M. Sydow, M. Pikua, and R. Schenkel, "The notion of diversity in graphical entity summarisation on semantic knowledge graphs," *Journal of Intelligent Information Systems*, vol. 41, no. 2, pp. 109–149, 2013.
[35] A. Bordes and E. Gabrilovich, "Constructing and mining web-scale knowledge graphs: Kdd 2014 tutorial," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1967–1967.
[36] N. Jayaram, A. Khan, C. Li, X. Yan, and R. Elmasri, "Querying knowledge graphs by example entity tuples," *arXiv*, 2013.
[37] H. Uszkoreit and F. Xu, "From strings to things sar-graphs: A new type of resource for connecting knowledge and language." in *NLP-DBPEDIA@ ISWC*, 2013.