

# TOWARDS EXPERTISE MODELLING FOR ROUTING DATA CLEANING TASKS WITHIN A COMMUNITY OF KNOWLEDGE WORKERS

(Research-in-Progress)  
(Data Scrubbing and Cleaning, Crowd Sourcing, Community Input)

**Umair ul Hassan**

Digital Enterprise Research Institute  
National University of Ireland  
Galway, Ireland  
umair.ul.hassan@deri.org

**Sean O’Riain**

Digital Enterprise Research Institute  
National University of Ireland  
Galway, Ireland  
sean.oriain@deri.org

**Edward Curry**

Digital Enterprise Research Institute  
National University of Ireland  
Galway, Ireland  
ed.curry@deri.org

**ABSTRACT:** Applications consuming data have to deal with variety of data quality issues such as missing values, duplication, incorrect values, etc. Although automatic approaches can be utilized for data cleaning the results can remain uncertain. Therefore updates suggested by automatic data cleaning algorithms require further human verification. This paper presents an approach for generating tasks for uncertain updates and routing these tasks to appropriate workers based on their expertise. Specifically the paper tackles the problem of modelling the expertise of knowledge workers for the purpose of routing tasks within collaborative data quality management. The proposed expertise model represents the profile of a worker against a set of concepts describing the data. A simple routing algorithm is employed for leveraging the expertise profiles for matching data cleaning tasks with workers. The proposed approach is evaluated on a real world dataset using human workers. The results demonstrate the effectiveness of using concepts for modelling expertise, in terms of likelihood of receiving responses to tasks routed to workers.

**Keywords:** data cleaning, crowd sourcing, web 2.0, linked data

## INTRODUCTION

The information systems of a business contain data on entities important to the business such as products, customers, suppliers, employees, etc. Entity information is spread across the organization, shared with partners, or even outside its boundaries of control, for example on the web. Maintaining a clean and consistent view of business critical entities is a core requirement of any knowledge based organization, as highlighted by a recent survey on the value of data analytics in organizations [1]. The study found that

more than 30% executives considered *integration, consistency, and trustworthiness* their top most data priorities. Most of the information quality research has focused on the development of sophisticated data quality tools and approaches such as Master Data Management. However these tools and techniques necessitate high technical expertise for successful implementation. Consequently, one of the major obstacles to data quality are the high operational costs due to limited availability of a few experts, and changes to business rules and policies [2], [3]. To overcome this limitation automatic or semi-automatic data cleaning algorithms can be used to improve data quality. However, the output of these algorithms can still require human review to ensure trust for decision making.

Involving the community of users in data management activities has shown promising results for maintaining high quality data [4]. Recent developments in *crowdsourcing* [5] and *human computation* [6] have fuelled the interest in algorithmic access to human workers, within or outside organizations, for performing computationally difficult tasks. Most of the current approaches of human computation publish tasks on task markets such as Amazon Mechanical Turk<sup>1</sup>. Therefore leaving the choice of task selection to the unknown workers, through search and/or browse capabilities of the platform. As a result the quality of responses provided by the workers may suffer from lack of domain knowledge or expertise for the task at hand. However, if the knowledge of workers' expertise is understood, tasks can be assigned to appropriate workers in a crowd or community. This process is known as *task routing*.

In this paper we propose a approach for task routing that profiles knowledge workers according to their expertise of concepts related to data quality issues and then assigns data quality tasks to appropriate workers. The approach is implemented in the *CAMEE* (Collaborative Management of Enterprise Entities) system. Given a set of data cleaning updates, *CAMEE* automatically converts them to feedback tasks for further verification from the group of knowledge workers considering their individual expertise levels. We argue that the expertise level of workers can be effectively measured against concepts associated with data quality tasks, where concepts are extracted from source data.

In this paper, we address the problem of building expertise profiles of worker and leveraging these profiles for routing tasks to appropriate workers. The contributions of this paper are as follows:

- An approach for modelling and assessment of knowledge worker's expertise with concepts and a prototype implementation of the approach using SKOS<sup>2</sup> concepts
- A simple concept matching approach for routing data quality tasks to appropriate worker
- A preliminary evaluation of proposed system on real world dataset with real world workers to demonstrate its effectiveness

The rest of this paper is organized as follows. Next section motivates the research work with respect to data quality management. Then we provide an overview of the system architecture and related research challenges. The implementation section details the prototype system using SKOS concepts for modelling expertise, as well as two approaches of building expertise model for task routing. The section on evaluation presents the experimental details and discusses the results. Finally we provide the review of existing work in closely related research areas and summarize the paper afterwards.

## MOTIVATION

*Master Data Management* (MDM) [7] has become a popular approach for managing quality of enterprise data. The main benefit of a successful MDM implementation is readily available high quality data about entities in an enterprise. Although attractive, recent studies estimate that more than 80% data integration

---

<sup>1</sup> <http://www.mturk.com>

<sup>2</sup> <http://www.w3.org/2004/02/skos/>

projects in enterprises either fail or overrun their budget [2], [8]. MDM is heavily centralized and labour intensive, where the cost and effort in terms of expertise can become prohibitively high. The main responsibility for data quality management lies with the MDM council in a *top-down* manner [9]. An MDM council usually includes members from senior management, business managers and data stewards.

The significant upfront costs in terms of development efforts and organizational changes make MDM difficult to implement successfully across large enterprises. The concentration of data management and stewardship between few highly skilled individuals, like developers and data experts, also proves to be a bottleneck. To this end, the lack of delegation of data management responsibilities is considered as one of most the significant barriers to data quality [2]. Due to the limited number of skilled human resources, only a small percentage of enterprise data comes under management. As a result, the scalability of MDM becomes a major issue when new sources of information are added over time. Not only are enterprises unable to cope with the scale of data generated within their boundaries. As the web data becomes important, there will be a need for enterprises to manage external data existing outside their boundaries within shared global information ecosystems [10].

Effectively involving a wider community of users within collaborative data cleaning and information management activities is attractive proposition. The *bottom-up* approach of involving crowds in creation and management of general knowledge has been demonstrated by projects like Freebase<sup>3</sup>, Wikipedia<sup>4</sup>, and DBpedia<sup>5</sup> [4]. Similarly data quality workload can be delegated to community of end-users by effectively guiding them towards specific tasks in *top-down* manner [11]. Sourcing data quality tasks to a community or crowd necessitates explicit control over the actions required from humans and their potential outcome.

*Human computation* [6] is a relatively recent field of research that focuses on the design of algorithms with operations or functions carried out by human workers. One of the major aspects of human computation is to understand the expertise of available humans and match them with the appropriate tasks. In this respect, systems using human computation need to overcome two challenges; 1) how to assess and model human expertise towards, and 2) how to effectively route tasks to appropriate workers. In this paper we outline a collaborative data quality management system that follows a human computation approach for involving end-users in the cleaning process. We introduce a concept based approach for modelling the expertise of human workers for task routing.

## CAMEE OVERVIEW

*CAMEE* follows a human computation approach that utilizes community participation to incrementally increase the quality of data. Using *CAMEE*, technical experts (e.g. developers, data stewards, and data analyst) define the data quality processes with the objective of routing tasks to human workers having relevant domain knowledge to complete the task. The worker may be employees of the organization or sourced from an online marketplace. The rest of this section describes the workflow of the system followed by discussion on challenges of expertise modelling and task routing.

### *System Workflow*

Figure 1 presents the high level workflow of the *CAMEE* system. The input to *CAMEE* is a dirty dataset that is assessed by *data cleaning algorithms* against pre-defined policies or rules, to identify data quality issues.

---

<sup>3</sup> <http://www.freebase.com>

<sup>4</sup> <http://www.wikipedia.org>

<sup>5</sup> <http://www.dbpedia.org>

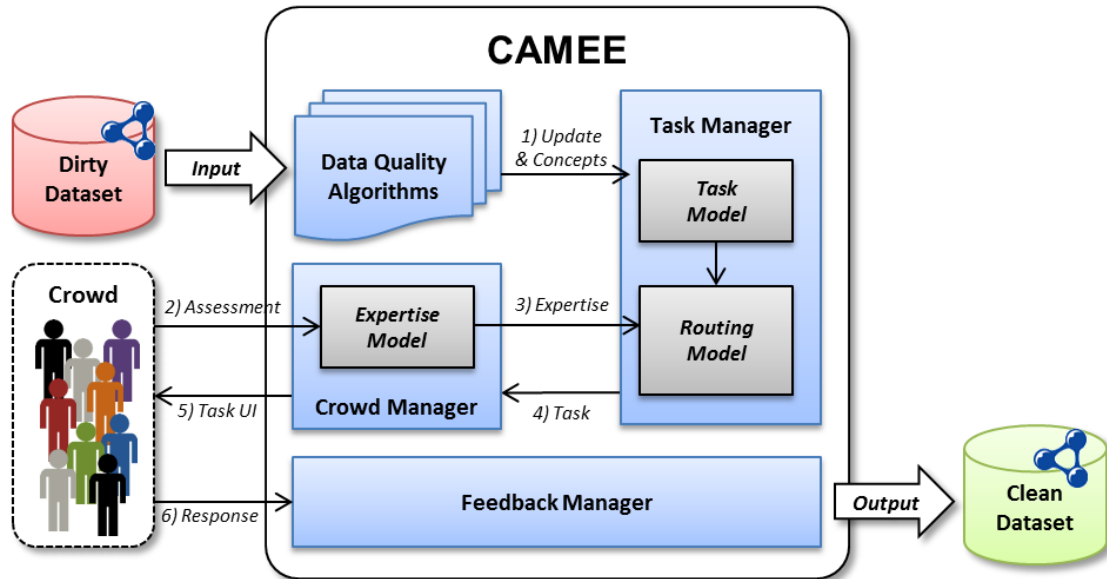


Figure 1: An example workflow of CAMEE for cleaning dataset with crowdsourcing.

- 1) Data quality algorithms suggest *updates* to the dataset for each data quality issue. The *concepts* describing the dataset are extracted and associated with each update. The suggested updates are fed to the *task manager* component, which converts an update into a task.
- 2) The *crowd manager* component maintains an *expertise model* by either soliciting expertise level directly from workers, or by calculating indirectly through their performance for test tasks with known responses.
- 3) The *routing model* matches each task with the appropriate worker according to their expertise, and then;
- 4) Submits the task to the crowd manager for execution.
- 5) The crowd manager renders each task using an appropriate user interface.
- 6) The *feedback manager* captures the response to the tasks and generates a cleaned dataset as output of the system.

### ***Expertise & Routing***

Human computation approaches rely on explicit control over routing of tasks to appropriate human workers. The tasks can be routed following a pull method by posting tasks on an online marketplace, such as Amazon Mechanical Turk. In pull method the decision of routing is delegated onto the humans themselves by allowing them to select tasks using search or browse features of the marketplace. On the other hand the push method of routing actively selects appropriate workers from a pool of available human resources. CAMEE follows push method of task routing that requires an understanding of the expertise of human workers for matching tasks to appropriate workers. The main challenges associated with push routing are

- How to represent domain knowledge of data quality task
- How to assess and represent expertise of workers for a particular domain of knowledge
- How to match domain of data quality task with expertise of workers

The expertise required to complete a data quality tasks not only depends on the type of task but also on the domain knowledge. In this paper we propose a concept based approach for addressing above mentioned challenges. We show that concepts extracted from the source data can be effectively used for

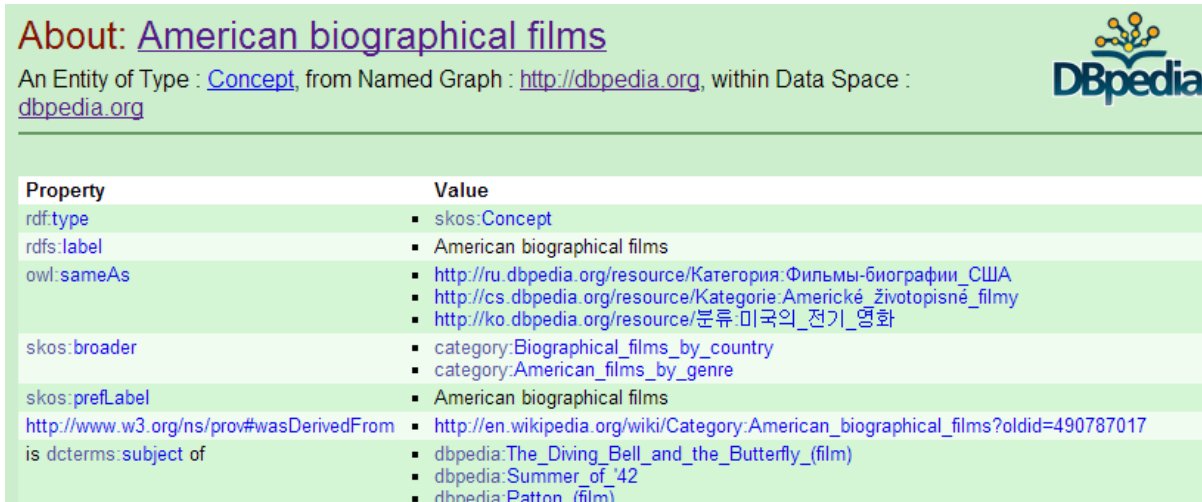
modelling worker expertise and routing tasks. In next section we describe an example implementation of the approach within CAMEE that exploits concepts in source data as the common denominator for annotating data quality tasks, building worker expertise, and routing tasks.

## CONCEPT-BASED EXPERTISE MODELLING WITHIN CAMEE

In this section we provide details of the prototype implementation of concept-based expertise modelling within CAMEE. We illustrate by example the application of concepts based expertise modelling and task routing within data quality management.

### SKOS Concepts

The *Simple Knowledge Organization System* (SKOS) is a W3C recommended data model designed to represent knowledge organization systems and share them through the Web [12]. The organization systems can include thesauri, subject headings, classification schemes, taxonomies, glossaries and other structured controlled vocabularies. In SKOS the basic element is a concept, identified by URI<sup>6</sup>, which is considered to be ‘unit of thought’; ideas, meanings or objects. Furthermore, SKOS defines attributes for labelling concepts with lexical strings and providing additional textual information regarding the concept. Concepts can be grouped into concept schemes and linked with other concepts by using semantic relationship hierarchical or associative attributes in SKOS. The overall objective of SKOS is to provide a common data model for knowledge organization systems, to facilitate their interoperability, as well as to make them machine-readable through a web-based data format called *Resource Description Framework*<sup>7</sup> (RDF). The usability of SKOS has been demonstrated with use cases of knowledge organization systems from life sciences, agriculture, product lifecycle, and media [13]. In this paper, we use the case of DBpedia [14] which is a structured knowledge base constructed by extracting and linking entities from Wikipedia. Figure 2 shows properties and values of concept *American\_biographical\_films* in DBpedia.



The screenshot shows the DBpedia interface for the concept 'American biographical films'. It includes the title 'About: American biographical films', the DBpedia logo, and a table of RDF properties and values. The table lists properties such as 'rdf:type', 'rdfs:label', 'owl:sameAs', 'skos:broader', 'skos:prefLabel', 'http://www.w3.org/ns/prov#wasDerivedFrom', and 'is dcterms:subject of', each with its corresponding value or list of values.

Property	Value
rdf:type	▪ skos:Concept
rdfs:label	▪ American biographical films
owl:sameAs	▪ <a href="http://ru.dbpedia.org/resource/Категория:Фильмы-биографии_США">http://ru.dbpedia.org/resource/Категория:Фильмы-биографии_США</a> ▪ <a href="http://cs.dbpedia.org/resource/Kategorie:Americké_životopisné_film_y">http://cs.dbpedia.org/resource/Kategorie:Americké_životopisné_film_y</a> ▪ <a href="http://ko.dbpedia.org/resource/분류:미국의_전기_영화">http://ko.dbpedia.org/resource/분류:미국의_전기_영화</a>
skos:broader	▪ category:Biographical_films_by_country ▪ category:American_films_by_genre
skos:prefLabel	▪ American biographical films
http://www.w3.org/ns/prov#wasDerivedFrom	▪ <a href="http://en.wikipedia.org/wiki/Category:American_biographical_films?oldid=490787017">http://en.wikipedia.org/wiki/Category:American_biographical_films?oldid=490787017</a>
is dcterms:subject of	▪ dbpedia:The_Diving_Bell_and_the_Butterfly_(film) ▪ dbpedia:Summer_of_'42 ▪ dbpedia:Patton_(film)

Figure 2: Screenshot of RDF data in DBpedia about the SKOS concept *American\_biographical\_films*

DBpedia converts Wikipedia articles to entities in RDF format through hand crafted mappings and natural language techniques. Similarly it converts concepts from Wikipedia category system to SKOS concepts. Figure 3 shows some attributes and concepts of the Wikipedia article for the movie “A Beautiful Mind” in RDF format.

<sup>6</sup> Uniform Resource Identifier

<sup>7</sup> <http://www.w3.org/RDF/>



Figure 3: Screenshot of RDF data in DBpedia about the movie “A Beautiful Mind”

In Figure 3 the *dbpedia-owl:starring* attribute have been extracted from the InfoBox of the Wikipedia article. The *dct:subject* attributes has been assigned the SKOS concept extracted from article’s categories box. For example, [http://dbpedia.org/resource/Category:American\\_biographical\\_films](http://dbpedia.org/resource/Category:American_biographical_films) represents the SKOS concept equivalent of Wikipedia category “American Biographical Films”. While the Wikipedia category system is collaboratively created and updated by editors, similar or even more sophisticated knowledge organization systems exists within large enterprises. There are tools<sup>8</sup> available for generation and management of SKOS concept schemes from existing taxonomies, vocabularies or knowledge organization systems. Figure 4 give an example use of SKOS concepts by CAMEE for representing domain of knowledge for data quality tasks, expertise of knowledge worker and task routing decisions.

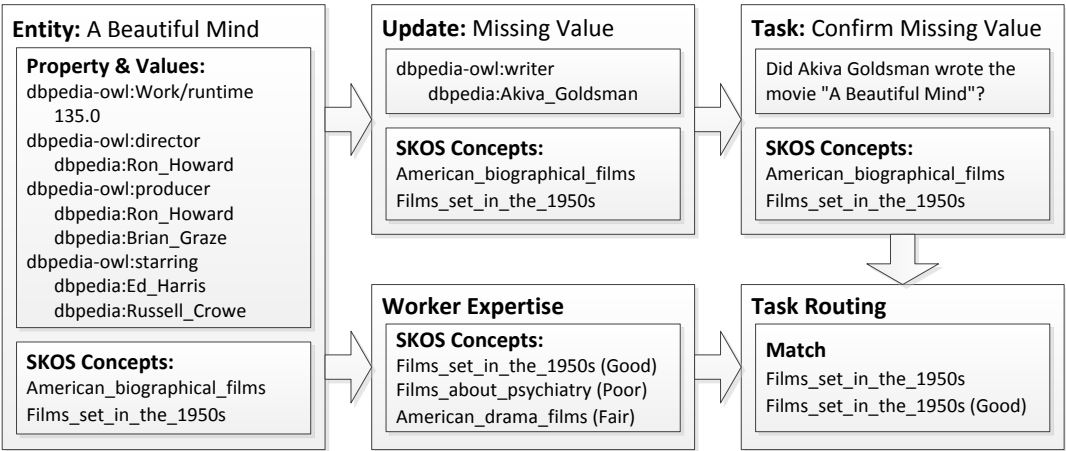


Figure 4: Example use of SKOS concepts for representing expertise and task routing in CAMEE

<sup>8</sup> <http://www.w3.org/2001/sw/wiki/SKOS>

## Expertise Modelling

SKOS provides a language to design knowledge structures in as simple as possible way. We use SKOS concepts, from source data, for modelling expertise requirements of tasks and knowledge level of workers for CAMEE. Assuming that the entities in the dataset have been annotated with some simple SKOS concept scheme as highlighted in Figure 2, the task manager associates concepts with the data quality task. For example the data quality task for the movie entity *A\_Beautiful\_Mind\_(film)* has *American\_biographical\_films*, *Best\_Drama\_Picture\_Golden\_Globe\_winners*, and *Films\_set\_in\_1950s* SKOS concepts associated with it. The crowd manager component builds worker profiles for the SKOS concepts according one of the following two approaches:

- *Self-Assessment (SA)*: In this approach a worker is asked to rate their knowledge level among the list of all concepts in the dataset.
- *Test Assessment (TA)*: A worker's knowledge expertise is based on her performance of data quality tasks with known answers, where each tasks has concepts associated with it.

For example, a worker can specify their knowledge level for *American\_biographical\_films* concepts as excellent for SA approach. However during the TA approach her responses for the test tasks associated with *American\_biographical\_films* can suggest a below average level of knowledge. Table 1 gives an example of expertise profiles for 3 workers on 4 concepts related to movies, where each value represents the knowledge level between the values of 0 and 1.

Concept	Worker 1	Worker 2	Worker 3
1990s_comedy-drama_films	0.6	0.2	0.2
Films_about_psychiatry	0.6	0.2	0.6
American_biographical_films	0.8	0.4	0.4
American_comedy-drama_films	0.8	0.6	0.6

Table 1: Example of matrix of expert profiles for 3 workers and 4 movie concepts

## Task Routing

The expertise model is exploited by the task routing model for matching tasks with appropriate knowledge workers. In this paper following matching strategies are employed for the purpose of routing

- *Random*: Sends a particular task to any randomly selected worker from the pool of all available workers. This routing strategy assumes unavailability of a worker's expertise model, thus serving as the baseline approach as well as fall back strategy.
- *Expertise Match*: This strategy ranks workers according to the weighted matching score between task concepts and the worker's expertise profile. The weights are based on the expertise model built earlier. The example task discussed would be routed to the worker with highest score for the *American\_biographical\_films*, *Films\_about\_psychiatry*, and *Films\_based\_on\_biographies* concepts

## EVALUATION

We performed an empirical evaluation of task routing based on the proposed expertise model using the two approaches; self-assessment and task-assessment. The two objectives of the experiments are 1) to

compare random routing without using workers’ expertise models versus routing based on matching task concepts and worker expertise, and 2) to investigate the best approach for building the worker expertise model. We evaluated if the concepts extracted from the dataset can be utilized effectively for representing the knowledge space of data quality tasks and worker expertise. In this regards we have explored the following proposition through empirical evaluation:

Data quality tasks routed using a concept-based expertise profiles have higher response rates if the expertise model is built using a task-assessment approach as compared to a self-assessment based approach.

## ***Experiments***

In this section we provide the details of the experiment design employed for the purpose of evaluation. We have divided the experimental evaluation in two stage process.

- The *assessment stage* focused on building the expertise model of workers. During this stage workers were asked to complete one assessment for each of the expertise building approaches. A simple 5 points belief scale (i.e. *none, poor, fair, good, and excellent*) was used for the self-assessment of knowledge about concepts. The workers were asked to provide responses to task-assessments based on Likert scale<sup>9</sup> (i.e. *don’t know, strongly disagree, disagree, neutral, agree, strongly agree*). *None* and *Don’t Know* were the default selected options for belief scale and Likert scale, respectively.
- The *routing stage* used the generated expertise for routing data quality tasks to appropriate knowledge workers. These responses to were used to calculate quality for final output dataset.

The response of workers for tasks routed to them is recorded against Likert scale with default response of “Don’t Know”. So for a particular approach a high percentage of workers providing “Don’t Know” responses indicate a low likeliness of getting data cleaned with help of workers. While a low percentage of “Don’t Know” responses indicated a high likeliness. In the rest of this section, we describe the datasets used for experiments, as well as the data quality tasks required to clean these datasets. Details of the population of knowledge workers and their characteristics are also discussed.

### **Dataset Description**

We have used a subset of DBpedia describing movies within the experimentation. A test dataset was created by selecting Academy Award and FilmFare Award winning movies, as well as the top 100 grossing movies from Hollywood and Bollywood. The DBpedia database provides variety of concept schemes for entities. However for the purpose of this experiment we selected 42 film genre concepts associated with movies. Detailed statistics of the dataset are listed in the Table 2.

<b>Characteristic</b>	<b>Value</b>
Number of entities (dbp:Film)	724
No. of concepts	42
No. of data quality tasks	230

Table 2: Characteristics of dataset describing award winning and top 100 grossing movies from Hollywood and Bollywood in DBpedia

<sup>9</sup> [http://en.wikipedia.org/wiki/Likert\\_scale](http://en.wikipedia.org/wiki/Likert_scale)



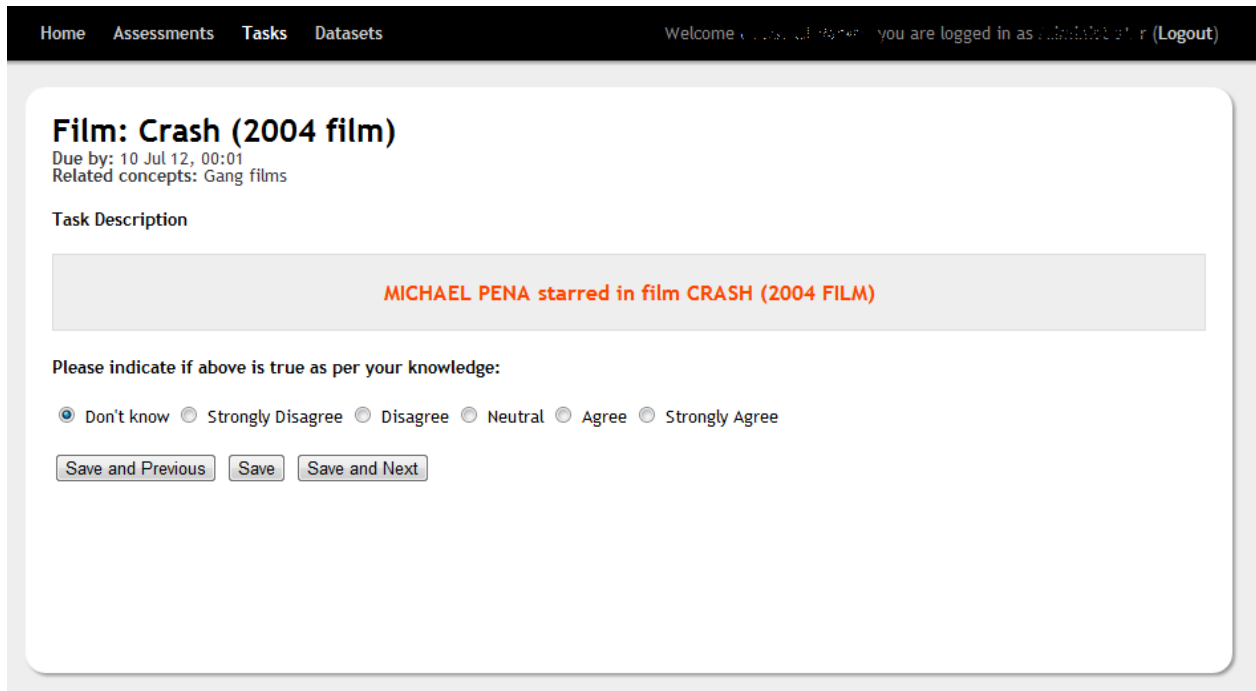
## Data Quality Tasks

The original movie dataset a variety of data quality issues. Table 3 highlights three particular types of issues. Each of these data quality issues is converted to a human computation task, which can be routed to knowledge workers. The conversion process involved creating a short question for the DQ issue, by using available data for the entity.

DQ Issue Type	Example question for DQ task
Identity Resolution	Does the following URIs represent the same entity? (Answer YES or NO) <i>http://dbpedia.org/resource/Shanghai_(2010_film)</i> <i>http://rdf.freebase.com/ns/m/047fjfr</i>
Missing Value	Did the following actor starred in the movie “Titanic”? (Answer YES or NO) <i>http://www.dbpedia.org/resource/bruce_willis</i>
Data Repair	Was the following movie released in 21-10-2011 or 21-10-2010? (Answer YES or NO) <i>http://www.dbpedia.org/resource/the_iron_lady</i>

Table 3: Examples of questions for the human computation tasks associated with specific data quality issues

The dataset was cleaned manually by an expert to serve as the gold standard. The data quality tasks were created by collecting correct and incorrect values for the “starring” attribute for movies. Figure 5 shows a screenshot of a human computation task.



Home Assessments Tasks Datasets Welcome, [User Name] you are logged in as [User Name] (Logout)

### Film: Crash (2004 film)

Due by: 10 Jul 12, 00:01  
Related concepts: Gang films

Task Description

**MICHAEL PENA starred in film CRASH (2004 FILM)**

Please indicate if above is true as per your knowledge:

Don't know  Strongly Disagree  Disagree  Neutral  Agree  Strongly Agree

Figure 5: Screenshot of the CAMEE prototype system for crowd sourcing data quality tasks

## Knowledge Workers

We recruited volunteer workers to perform the human computation tasks for data quality. The final community of workers contained people from 3 regions of worlds (Europe, South Asia, and Middle East) having varying knowledge about the movie dataset, as shown in Table 4.

Characteristic	Value
No. of Workers	11
Tasks for Assessment Stage	100
Tasks for Routing Stage	130

Table 4: Characteristics of knowledge worker recruited for the experiments, as well as statistics of tasks assigned to them during test stage

## Results

The following results show the distribution of responses for the *Random* routing as compared to *Expertise Match* based routing coupled with the expertise modelling approaches. As expected both matching based routing strategies outperform random routing of tasks. The data confirms that building expertise models based on performance on task-assessments is a better approach as compared just soliciting self-assessment of knowledge about concepts.

Expertise Approach	Random	Self-Assessment + Matching	Task Assessment + Matching
Don't know	73.85%	56.15%	36.92%
Strongly Disagree	6.92%	14.62%	16.15%
Disagree	6.15%	5.38%	13.08%
Neutral	0.00%	3.85%	7.69%
Agree	3.08%	5.38%	8.46%
Strongly Agree	10.00%	14.62%	17.69%

Table 5: Distribution of responses during routing stage, for 3 task routing approaches. A high percentage of “Don’t Know” response indicates that the tasks has been routed to worker with no domain knowledge.

## RELATED WORK

The crowdsourcing approaches for data management activities can be categorized into three approaches; *algorithmic approaches*, *crowd-sourced databases* and *application platforms*.

*Algorithmic approaches* focus on the designing algorithms for reducing uncertainty of data management with human computed functions. In these approaches human attention is utilized to support data management system in different activities, such as schema matching [15], entity resolution [16] and data repair [17]. The objective of algorithmic approaches is to help increase utility of human attention through optimization of specific data management activities. Consequently the evaluation of these approaches focus on the measurement of incremental utility improvement after successive human interventions. Our work focuses on modelling expertise required for data quality tasks and building worker profiles to facilitate task routing.

*Crowd-sourced database* systems focus on providing programmatic access to human computation platforms for database operations such as joins, sorts, and inserts. This facilitates platform independence

with respect to the details of access to human services. Typically existing query languages are extended to minimize the learning curve associated with programming human computation. For example, CrowdDB [18] extends *standard query language* to provide database services on top on crowd sourcing platforms. An initial list of information quality problems which can be solved with crowdsourcing have been identified in [19]. The application of human computation has been demonstrated for data management problems such as *data ranking* [20], *relevance assessment* [21] and *entity linking* [22]. These research efforts focus on improving the quality of crowd responses through various task aggregation techniques after execution. Instead we focus the step before execution of tasks; improving the routing of tasks to workers with appropriate domain knowledge and expertise.

*Application platforms* extend existing applications with custom human computation capabilities, thus enabling crowd services in applications. These approaches do not depend on external platforms for human services as compared to previous categories. Freebase supported by a human computation platform called RABj [23], which allows users to distribute specific tasks to communities of paid or volunteering worker. Similarly, MOBS [24] provides a tool extension approach for enabling crowd sourcing of schema matching applications. Both RABj and MOBS are crowd sourcing platforms tailored for specific data management applications. We propose CAMEE; a human computation based approach for guided data cleaning. The objective of CAMEE is to facilitate task routing for effective utilization of human attention in collaborative data cleaning processes.

*Expert finding* has been the subject of a considerable amount of research in the Information Retrieval community [25]. The expert finding problem involves ranking the list of experts according to their knowledge about a given topic or query. Generally, some web-based or enterprise text corpus is utilized to uncover associations between experts and topics [26]. On the other hand, *expert profiling* is defined as the opposite process of determining the list of topics that an expert is knowledge about [27]. In both cases, current approaches mine existing text corpus to determine worker and topics associations. By contrast, in this paper we are interested in profiling expertise of workers for finding task and worker associations. We cast this problem in a data cleaning scenario where we building profiles by only using source data. We assume that the source data does not provide any evidence of worker expertise in form of person and topic associations. Instead we demonstrate the effective use of SKOS for the purpose of expertise profiling and task routing with in data cleaning scenario.

## **SUMMARY AND FUTURE WORK**

This paper presents an concepts based approach for routing data quality tasks to appropriate workers based on an their knowledge and expertise. An expertise model for representing worker profiles against a set of concepts from the dataset is described. The approach is validated with a simple routing algorithm for exploiting expertise model based on either concept selection or task performance. The approach is evaluated on real world datasets using human workers. The results demonstrate the effectiveness of using concept based profiles for soliciting higher number of responses from workers

In this paper we described the architecture of CAMEE and its use of SKOS concepts for modelling expertise for tasks and knowledge worker. As the part of future work we plan to expand our analysis of the system to effect of various expertise assessment methods and task routing methods on quality of task routing. Further research is also required into the effective balancing of the community workload under constraints such as cost, latency, and motivation. We plan to investigate the utility of CAMEE in real world information management scenario that deals with multiple data sources and heterogeneity problems, such as enterprise energy management [28].

## ACKNOWLEDGEMENTS

We thank all the volunteers, and all publications support and staff, who wrote and provided helpful comments on previous versions of this document. Some of the references cited in this paper are included for illustrative purposes only. The work presented in this paper is funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion- 2).

## REFERENCES

- [1] S. Lavalle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, "Big Data, Analytics and the Path from Insights to Value," *MIT Sloan Management Review*, vol. 52, no. 2, pp. 21–32, 2011.
- [2] A. Haug and J. S. Arlbjørn, "Barriers to master data quality," *Journal of Enterprise Information Management*, vol. 24, no. 3, pp. 288–303, 2011.
- [3] R. Silvola, O. Jaaskelainen, H. Kropsu-Vehkaperä, and H. Haapasalo, "Managing one master data – challenges and preconditions," *Industrial Management & Data Systems*, vol. 111, no. 1, pp. 146–162, 2011.
- [4] E. Curry, A. Freitas, and S. O. Riain, "The Role of Community-Driven Data Curation for Enterprises," in *Linking Enterprise Data*, D. Wood, Ed. Boston, MA: Springer US, 2010, pp. 25–47.
- [5] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the World-Wide Web," *Communications of the ACM*, vol. 54, no. 4, p. 86, Apr. 2011.
- [6] E. Law and L. von Ahn, "Human Computation," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 5, no. 3, pp. 1–121, Jun. 2011.
- [7] D. Loshin, *Master Data Management*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008.
- [8] B. Otto and A. Reichert, "Organizing Master Data Management: Findings from an Expert Survey," in *Proceedings of the 2010 ACM Symposium on Applied Computing - SAC '10*, 2010, pp. 106–110.
- [9] K. Weber, B. Otto, and H. Österle, "One Size Does Not Fit All---A Contingency Approach to Data Governance," *Journal of Data and Information Quality*, vol. 1, no. 1, pp. 1–27, Jun. 2009.
- [10] S. O’Riain, E. Curry, and A. Harth, "XBRL and open data for global financial ecosystems: A linked data approach," *International Journal of Accounting Information Systems*, Mar. 2012.
- [11] U. Ul Hassan, S. O’Riain, and E. Curry, "Leveraging Matching Dependencies for Guided User Feedback in Linked Data Applications," in *9th International Workshop on Information Integration on the Web IIWeb2012*, 2012.
- [12] A. Miles and J. R. Pérez-Agüera, "SKOS: Simple Knowledge Organisation for the Web," *Cataloging & Classification Quarterly*, vol. 43, no. 3–4, pp. 69–83, Apr. 2007.
- [13] A. Isaac, J. Phipps, and D. Rubin, "SKOS Use Cases and Requirements." [Online]. Available: <http://www.w3.org/TR/skos-ucr/>. [Accessed: 28-Sep-2012].
- [14] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia - A crystallization point for the Web of Data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, pp. 154–165, Sep. 2009.
- [15] K. Belhajjame, N. W. Paton, S. M. Embury, A. A. A. Fernandes, and C. Hedeler, "Feedback-based annotation, selection and refinement of schema mappings for dataspace," in *Proceedings of the 13th International Conference on Extending Database Technology - EDBT '10*, 2010, p. 573.
- [16] S. R. Jeffery, M. J. Franklin, and A. Y. Halevy, "Pay-as-you-go user feedback for dataspace systems," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data - SIGMOD '08*, 2008, pp. 847–860.
- [17] M. Yakout, A. K. Elmagarmid, J. Neville, M. Ouzzani, and I. F. Ilyas, "Guided Data Repair," *Proceedings of the VLDB Endowment*, vol. 4, no. 5, pp. 279–289, 2011.
- [18] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin, "CrowdDB: Answering Queries with Crowdsourcing," in *Proceedings of the 2011 international conference on Management of data - SIGMOD '11*, 2011, p. 61.
- [19] P. Wichmann, A. Borek, R. Kern, P. Woodall, A. K. Parlikad, and G. Satzger, "Exploring the 'Crowd' as Enabler of Better Information Quality," in *Proceedings of the 16th International Conference on Information Quality*, 2011, pp. 302–312.
- [20] A. Marcus, E. Wu, D. Karger, S. Madden, and R. Miller, "Human-powered Sorts and Joins," *Proceedings of VLDB Endowment*, vol. 5, no. 1, 2012.

- [21] C. Grady and M. Lease, "Crowdsourcing document relevance assessment with Mechanical Turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 172–179.
- [22] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux, "ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking," in *Proceedings of the 21st international conference on World Wide Web - WWW '12*, 2012, p. 469.
- [23] S. Kochhar, S. Mazzocchi, and P. Paritosh, "The anatomy of a large-scale human computation engine," in *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*, 2010, pp. 10–17.
- [24] R. McCann, W. Shen, and A. Doan, "Matching Schemas in Online Communities: A Web 2.0 Approach," in *2008 IEEE 24th International Conference on Data Engineering*, 2008, vol. 00, pp. 110–119.
- [25] K. Balog, L. Azzopardi, and M. de Rijke, "Formal models for expert finding in enterprise corpora," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, 2006, p. 43.
- [26] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch, "Broad expertise retrieval in sparse data environments," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, 2007, p. 551.
- [27] K. Balog and M. De Rijke, "Determining expert profiles (with an application to expert finding)," in *Proceedings of the 20th international joint conference on Artificial intelligence*, 2007, pp. 2657–2662.
- [28] E. Curry, S. Hasan, and S. O'Riain, "Enterprise Energy Management using a Linked Dataspace for Energy Intelligence," in *Second IFIP Conference on Sustainable Internet and ICT for Sustainability*, 2012.