# Distributed Human Computation Framework for Linked Data Co-reference Resolution

Yang Yang[1], Priyanka Singh[1], Jiadi Yao[1], Ching-man Au Yeung[2],
Amir Zareian[1], Xiaowei Wang[1], Zhonglun Cai[1], Manuel Salvadores[1], Nicholas Gibbins[1], Wendy Hall[1], and Nigel Shadbolt[1]

[1] Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science
University of Southampton, UK
[2] NTT Communication Science Laboratories
2-4 Hikaridai Seika-cho Soraku-gun
Kyoto, 619-0237, Japan
{yang.yang,ps1w07,jy2e08,ms8,nmg,wh,nrs}@.soton.ac.uk
auyeung@cslab.kecl.ntt.co.jp

**Abstract.** Distributed Human Computation (DHC) is used to solve computational problems by incorporating the collaborative effort of a large number of humans. It is also a solution to AI-complete problems such as natural language processing. The Semantic Web with its root in AI has many research problems that are considered as AI-complete. E.g. co-reference resolution, which involves determining whether different URIs refer to the same entity, is a significant hurdle to overcome in the realisation of large-scale Semantic Web applications. In this paper, we propose a framework for building a DHC system on top of the Linked Data Cloud to solve various computational problems. To demonstrate the concept, we are focusing on handling the co-reference resolution when integrating distributed datasets. Traditionally machine-learning algorithms are used as a solution for this but they are often computationally expensive, error-prone and do not scale. We designed a DHC system named iamResearcher, which solves the scientific publication author identity co-reference problem when integrating distributed bibliographic datasets. In our system, we aggregated 6 million bibliographic data from various publication repositories. Users can sign up to the system to audit and align their own publications, thus solving the co-reference problem in a distributed manner. The aggregated results are dereferenceable in the Open Linked Data Cloud.

## 1 Introduction

AI-complete problem is a set of problems found in areas such as image analysis, speech recognition and natural language processing that is difficult for computers to solve effectively but they are relatively easy tasks for humans [14]. Distributed Human Computation (DHC) [11] systems are designed to solve this kind of problems by incorporating collaborative efforts from a large number of humans. This approach is also known as crowdsourcing with computational purpose and in the Web 2.0 term, it's referred as participatory or social systems. For instance, reCAPTCHAs [17] is widely used on the Web to aid transcribing texts of old books that cannot be automatically processed by optical character recognition systems. The Semantic Web is envisioned to be a decentralised worldwide information space for sharing machine-readable data with a minimal cost of integration overheads [13]. However, there are many challenging research problems in the Semantic Web that are considered to be AI-complete, such as co-reference resolution, i.e. determining whether different URIs refer to the same identity [7].

In the recent years, there is an increasing number of linked datasets available on the Web. However, cross-reference and linkage between datasets are sparse as they cannot be easily created automatically. When creating a link between two datasets, intuitively we would consider linking the data that refer to the same thing as a bridge between the two. For instance, DBpedia has a URI referring to one of our authors, Nigel Shadbolt. This can be linked to the URI referring to N. Shadbolt in the Eprints repository dataset because they refer to the same person. Users can then follow the DBpedia URI and find out more about this person's publications. Various machine learning and heuristic algorithms have been proposed to automatically solve this co-referencing problem. However, these approaches are often computationally expensive, error-prone, require some training data, or are difficult to deploy on a large scale.

In this paper, we propose the idea of combining DHC system with Linked Data to create an ecosystem to solve computational problems and facilitate the deployment of Semantic Web. To demonstrate the concept, we focus on the design of a DHC system, iamResearcher [3] that aims to solve the co-referencing problem using DHC.

## 2 Background

### 2.1 Co-reference Resolution in the Semantic Web

There are many traditional approaches to perform co-reference resolution on the Web. Besides various natural language processing and machine learning techniques, there are also co-reference resolution systems that are especially designed to use in the Semantic Web to resolve URIs and name ambiguities.

In the area of machine learning Soon et al [16] resolved noun phrases by creating a co-reference relation and measuring the distance between two identities in order to find matches between nouns. Ng et al. [10] improved their algorithm by including more sophisticated linguistic knowledge to improve precision. In both cases, the authors found that performance dropped significantly when the dataset became larger and human intervention was required to solve co-references that were not accurately resolved automatically. Regarding author

[3] `http://www.iamresearcher.soton.ac.uk/` for University of Southampton members access only, and `http://www.iamreseacrher.com` for Global users

name ambiguities, Kang et al [8] had shown that co-authorship is a very reliable and decisive method to validate the identity of an author when there were namesakes. They proposed that author name disambiguation can be solved by clustering similar names into groups of identities and making use of other available information such as email addresses and publication titles to resolve the issue.

While, in an ideal Semantic Web, the identity of one person may be represented by different URIs in different systems. Sleeman et al. [15] proposed to use a rules-based model and a vector space model to cluster entities into groups of co-references. Whereas, Glaser et al [4] proposed the Co-reference Resolution Service to facilitate management of URI co-references. Salvadores et al [12] used LinksB2N algorithm to discover overlapping RDF data repositories to integrate datasets using clustering technique to find equivalent data.

These methods somewhat solves the co-reference problem but Semantic Web contains many highly complex data and these algorithms are insufficient in addressing the distinction between two URIs when they represent different entities in different context.

### 2.2 Human Computation

Human computation is a method of making use of the collaborative effort of a large number of humans to solve problems that are still difficult for computers to perform accurately. These tasks include natural language processing, speech recognition and image processing, which are relatively easy for human beings. Nowadays, people are more engaged into social activities on the Web, they collaborate and share information with one another, Wikipedia and Twitter are some of the many examples. The combination of the Social Web and human computation provides many opportunities to solve difficult computational problems.

reCAPTCHA, a system for distinguishing between humans and automated computer bots on the Web and at the same time it helps the digitization of millions of books [17] is a popular example of DHC. It proves that when a proper monitoring process is available and when users have the motivation or incentive to use such a system, one can collect reliable information for solving difficult computational problems. Albors et al. [1] discussed about the evolution of information through the effort of crowdsourcing. They mentioned Wikipedia and folksonomies as examples, where users are both the creators and consumers of the shared data, thus creating an ecosystem for the growth and development of information that ultimately benefit the users themselves. Gruber [6] discussed the structure of a collective knowledge system in which users interacted with one another and contributed their knowledge, while machines facilitated communications and retrieval of resources in the background, aggregating the contributions of individual users.

The following section discusses the implementation of DHC framework to solve the co-reference resolution in our system.

## 3 Linked Data Ecosystem Framework

At present there are 203 RDF datasets that have been published on the Linked Data Cloud.[4] Although this is encouraging, we are still relatively far from the Semantic Web envisioned by Tim Berners-Lee [2]. There are still many challenging research questions that are needed to be solved. From our experience in carrying out the Enakting project,[5] whose goal is to build the pragmatic Semantic Web, we have identified several challenges in Linked Data, such as co-reference resolution, ontology mapping, resource discovery, and federated query from multiple datasets [9]. Many research efforts in the past have been devoted to develop heuristic machine learning algorithm to solve these problems. However, these automated solutions do not necessarily solve these problems accurately.

Here, we propose to solve these problems by using DHC approach to build linked data ecosystem in which difficult computational tasks are distributed to the users in the system. And by ecosystem we mean that it is a self-sufficient system that can provide a long-term solution to a particular problem. For instance, an automated Semantic Web reasoner is likely to fail to return an answer when querying incomplete or noisy data. One can imagine a DHC system that can overcome this problem by enabling distributed reasoning on a subset of data with facilitation from human in certain decision making processes.

By studying different DHC systems, we have identified a list of common characteristics and designed a Linked Data Ecosystem Framework as depicted in Figure 1. To design an ecosystem first we need to identify the system stakeholders, i.e. the target data consumers and publishers. Next, we have the four major components for sustainability, namely incentive, human interface, data aggregation and quality control.

**Incentive.** We need to make sure that users have the incentive to use the system and therefore contribute to solving the problem which can manifest in different modality in different system. For instance, users want to use a system because they get paid, gain reputation, or simply because it is fun to use. This requires anthropological studies of the system stakeholders and we can design the system based on analysis of the generic usefulness of the system for the targeted crowd.

**Human Interface to solve computational problem.** This is the core of the system. It requires an interface that is applicable to the individual or small group of people to solve a computational problem in a distributed manner. For many problems, the system can use heuristic method to automate certain tasks to assist the human contributors.

**Aggregation.** The system combines the distributed human computation and heuristic algorithm output and aggregate the results to solve the global problem.

**Quality Control.** How does the system cope with possibility of fraud or incorrect answers to ensure some level of quality in the solution to the overall

[4] `http://www4.wiwiss.fu-berlin.de/lodcloud/state/` as on 22nd September 2010

[5] `http://www.enakting.org`

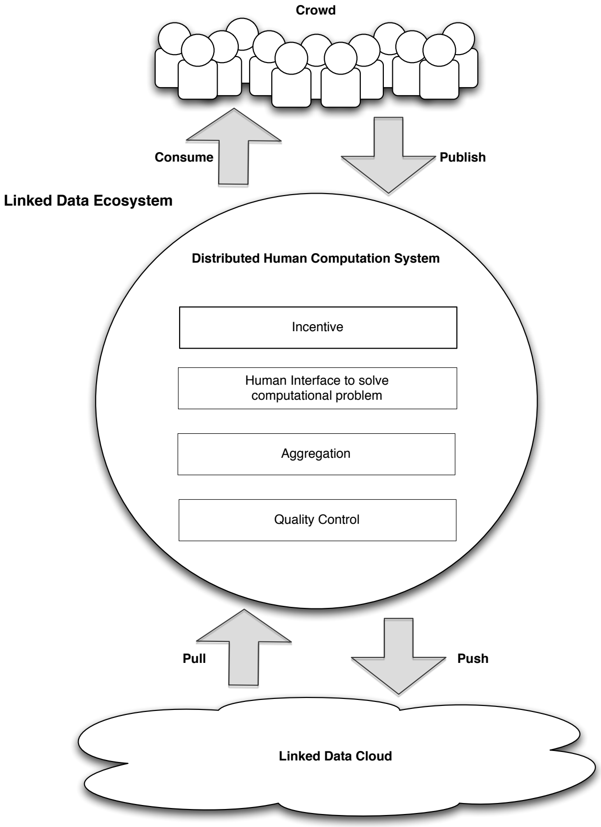


**Fig. 1.** Linked Data Ecosystem Framework

problems? The quality control in this framework acts as a layer to ensure the quality of the data to be pushed into the Linked Data Cloud.

In the following section, we apply the framework to a specific scenario–solving the co-reference problem in linked data.

## 4 Designing iamResearcher

The co-reference problem we are trying to solve in this paper is the name ambiguity problem in distributed digital libraries. Here is a typical scenario. In the Eprints[6] repository we have a list of publications authored by a person named Wendy Hall. In the PubMed[7] repository we have another list of publications

[6] http://www.eprints.org/

[7] http://www.ncbi.nlm.nih.gov/pubmed

authored by a person named W. Hall. If we want to design an expert finder algorithm that can rank researchers' expertise based on their publications, we must decide whether these two names refer to the same person.

Most of the large scale digital library repositories nowadays are not capable of resolving co-referencing and ambiguities. This is because it is difficult to determine if W. Hall is Wendy Hall or William Hall. Names of researchers are usually incomplete or inconsistent across different digital libraries. In particular, the name can be written with different combinations of the initials, the first name, middle name and last name. There can even be incorrect spellings. For example, within our own institutional Eprints repository, there can be as many as six different ways of naming any individual author. The extent of this name ambiguity can be seen within the UK research community based on the analysis of the Research Assessment Exercise 2001 records we did in the previous AKT project.[8] Within the list of researcher names in the institutional submissions, 10% of names lead to clashes between two or more individuals. If the names are restricted to a single initial, the proportion of clashes rises to 17% [5]. This situation can be more severe on the global scale. The VIAF project [9] also designed a service to integrate different global libraries using a heuristic name-matching algorithm in bibliographic record clustering allowing national and regional variation which is difficult to make an alignment.

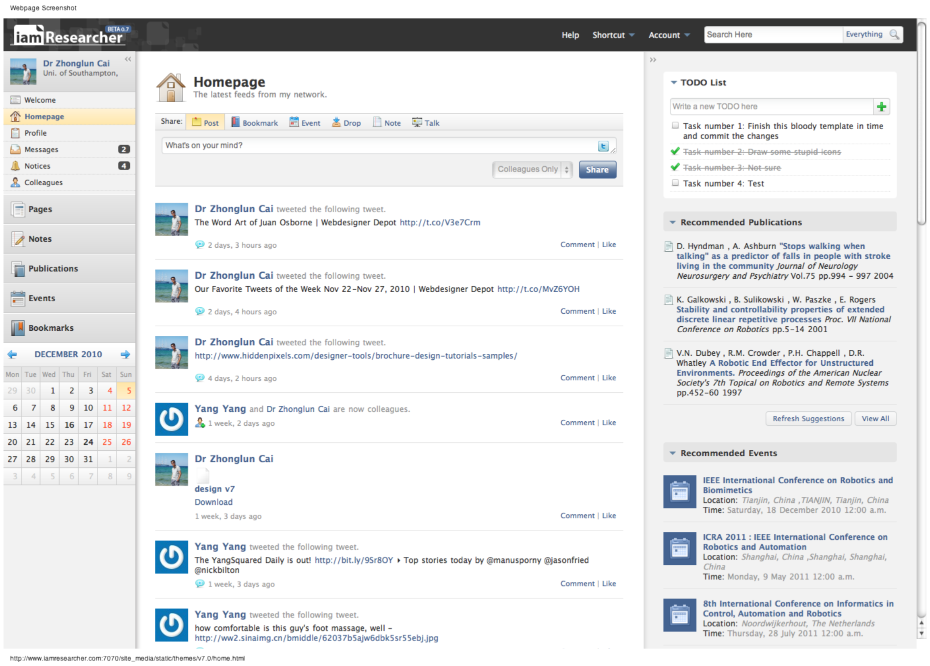
Co-reference problem has been well studied in computational linguistics. How do we determine if two things are the same? Leibniz's Law [3] states that 'X is the same as Y if, and only if X and Y have all the same properties and relations; thus, whatever is true of X is also true of Y, and vice-versa'. Based on this concept, we can compare the identities' relations and properties to determine if they are the same. For instance, we can check whether two names have the same affiliation and the same email address. However, in the real world, different information can be missing in different publication repositories. Even when all the information is available for comparison, one still have to consider the fact that properties of the same person can change over time. For example, when a researcher moves from one institution to another, his/her email address is likely to change.

As mentioned before, in order to derive the correct interpretation of a name, it should be connected to the right individuals. Therefore we propose to link the publication data with its individual author to solve the name ambiguity problem as the author would have the best knowledge about their own publication. Therefore, the fundamental ideas is that we aggregate bibliographic data from various repositories and ask users to audit the data and make alignment with their own publication data. This solves the name ambiguity co-reference problem.

Applying our framework, first we need to identify the stakeholders in our system–the data publishers and consumers. In our case, researchers play both roles. With above requirements, we designed a system that link all researchers and publications together. The system automatically pulls out all publications

[8] http://www.aktors.org/akt/

[9] http://www.viaf.org



Dr Zhonglun Cai
Uni. of Southampton,

DECEMBER 2010

| Mon | Tue | Wed | Thu | Fri | Sat | Sun |
|---|---|---|---|---|---|---|
| 29 | 30 | 1 | 2 | 3 | 4 | 5 |
| 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| 20 | 21 | 22 | 23 | 24 | 25 | 26 |
| 27 | 28 | 29 | 30 | 31 | 1 | 2 |
| 3 | 4 | 5 | 6 | 7 | 8 | 9 |

Homepage
The latest feeds from my network.

Dr Zhonglun Cai tweeted the following tweet.
The Word Art of Juan Osborne | Webdesigner Depot http://t.co/V3e7Crm
2 days, 3 hours ago Comment | Like

Dr Zhonglun Cai tweeted the following tweet.
Our Favorite Tweets of the Week Nov 22–Nov 27, 2010 | Webdesigner Depot http://t.co/MvZ6YOH
2 days, 4 hours ago Comment | Like

Dr Zhonglun Cai tweeted the following tweet.
http://www.hiddenpixels.com/designer-tools/brochure-design-tutorials-samples/
4 days, 2 hours ago Comment | Like

Yang Yang and Dr Zhonglun Cai are now colleagues.
1 week, 2 days ago Comment | Like

Dr Zhonglun Cai
design v7
Download
1 week, 3 days ago Comment | Like

Yang Yang tweeted the following tweet.
The YangSquared Daily is out! http://bit.ly/9Sr8OY ▸ Top stories today by @manusporny @jasonfried @nickbilton
1 week, 3 days ago Comment | Like

Yang Yang tweeted the following tweet.
how comfortable is this guy's foot massage, well – http://ww2.sinaimg.cn/bmiddle/62037b5ajw6dbk5sr55ebj.jpg

TODO List

- ☐ Task number 1: Finish this bloody template in time and commit the changes
- ✓ ~~Task number 2: Draw some stupid icons~~
- ✓ ~~Task number 3: Not sure~~
- ☐ Task number 4: Test

Recommended Publications

D. Hyndman , A. Ashburn "Stops walking when talking" as a predictor of falls in people with stroke living in the community *Journal of Neurology Neurosurgery and Psychiatry* Vol.75 pp.994 – 997 2004

K. Galkowski , B. Sulikowski , W. Paszke , E. Rogers Stability and controllability properties of extended discrete linear repetitive processes *Proc. VII National Conference on Robotics* pp.5–14 2001

V.N. Dubey , R.M. Crowder , P.H. Chappell , D.R. Whatley A Robotic End Effector for Unstructured Environments. *Proceedings of the American Nuclear Society's 7th Topical on Robotics and Remote Systems* pp.452–60 1997

Recommended Events

IEEE International Conference on Robotics and Biomimetics
Location: *Tianjin, China ,TIANJIN, Tianjin, China*
Time: Saturday, 18 December 2010 12:00 a.m.

ICRA 2011 : IEEE International Conference on Robotics and Automation
Location: *Shanghai, China ,Shanghai, Shanghai, China*
Time: Monday, 9 May 2011 12:00 a.m.

8th International Conference on Informatics in Control, Automation and Robotics
Location: *Noordwijkerhout, The Netherlands*
Time: Thursday, 28 July 2011 12:00 a.m.

**Fig. 2.** iamResearcher User Homepage

from various resources (as mentioned in Table 1) for researchers to audit and align the data. By analysing the network graph, researchers are then linked to each other via the co-authorships of the publications. The co-authorship often reflects their professional social network - if you often write papers with certain set of authors, most likely they are your colleagues. Based on this we designed a professional network portal-like application [18] - Researchers signup on the system to find their publications and establish the colleagueship with their co-authors and so on.

The general incentive for data consumer to use the system is that they can find experts and publications in their research field. The general incentive for the data publishers to use the system is that by creating their own list of publications they enable other researcher to find, read and cite their work. A researcher's scientific publication can evidently reflect his/her expertise. Therefore, individual may also be motivated to set up a list of publication for this purpose as well. To amplify the usage of the system, we also designed list of generic researcher oriented services like publication and research events recommendation based on their research interests, easy communication with their colleagues, group management system, bookmark management system etc. to encourage researcher to collaborate and use the system on the daily basis. We also make the user's FOAF profile with their publications dereferenceable in the Open Linked Data

Cloud. Figure 2 illustrates the homepage of the system showing status updates from their colleagues and recommended publications and conferences.

In the following section, we will elaborate the system design of the co-reference management and how we deal with quality control issues.

## 5 Co-reference Resolution

We have harvested metadata of publications from various repositories and databases. Table 1 gives an overview of the data we have collected.

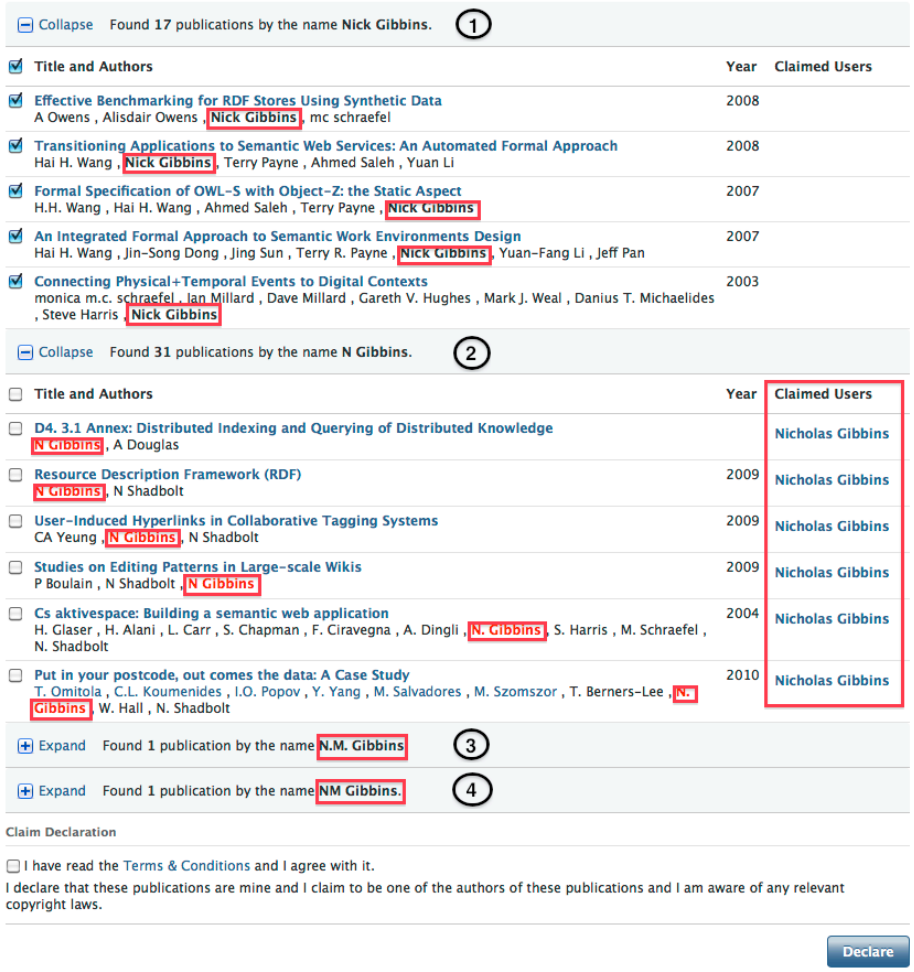
| **Source** | **Subjects Cover** | **Paper's Source** | **Papers Extracted** |
|---|---|---|---:|
| PubMed | Life sciences, Medicine | Peer-reviewed journal articles | 1381081 |
| Institutional EPrints | Multi-discipline | Preprint papers uploaded by researchers from each institute | 203387 |
| arXiv | Mathematics, Physics and Biology | Preprint papers uploaded by researchers | 478092 |
| DBLP | Computer science | Papers harvested from VLDB, IEEE, ACM | 1394314 |
| Econpapers | Economics | Part of RePEc | 361224 |
| Citeseer | Information Sciences, Engineering Sciences | Papers harvest from the web according to rules | 345821 |
| PANGAEA | Geoscientific and Environmental Sciences | Data submitted by researchers across the world | 576939 |
| Others | Multi-discipline | Papers harvested from search engine and numerous databases | 213276 |

**Table 1.** Dataset Source

Our co-reference system is designed as a two-stage process. Firstly, we used heuristic name matching algorithms to pull out all the possible combination and spelling of authors' names. Secondly, we let users audit the data by allowing them to select the publications from the resulted list.

When users register an account, we ask for their first and last name, our interface clearly states not to enter fake names or aliases as the system use their names to search for their publications and an incorrect name would lead the system in getting no matches or wrong matches.

The name-matching algorithm performs three types of matches: full name match, exact initial match and loose initial match.



⊟ Collapse Found **17** publications by the name **Nick Gibbins.**

| ☑ Title and Authors | Year | Claimed Users |
|---|---|---|
| ☑ Effective Benchmarking for RDF Stores Using Synthetic Data<br>A Owens , Alisdair Owens , Nick Gibbins , mc schraefel | 2008 | |
| ☑ Transitioning Applications to Semantic Web Services: An Automated Formal Approach<br>Hai H. Wang , Nick Gibbins , Terry Payne , Ahmed Saleh , Yuan Li | 2008 | |
| ☑ Formal Specification of OWL-S with Object-Z: the Static Aspect<br>H.H. Wang , Hai H. Wang , Ahmed Saleh , Terry Payne , Nick Gibbins | 2007 | |
| ☑ An Integrated Formal Approach to Semantic Work Environments Design<br>Hai H. Wang , Jin-Song Dong , Jing Sun , Terry R. Payne , Nick Gibbins , Yuan-Fang Li , Jeff Pan | 2007 | |
| ☑ Connecting Physical+Temporal Events to Digital Contexts<br>monica m.c. schraefel , Ian Millard , Dave Millard , Gareth V. Hughes , Mark J. Weal , Danius T. Michaelides , Steve Harris , Nick Gibbins | 2003 | |

⊟ Collapse Found **31** publications by the name **N Gibbins.**

| □ Title and Authors | Year | Claimed Users |
|---|---|---|
| □ D4. 3.1 Annex: Distributed Indexing and Querying of Distributed Knowledge<br>N Gibbins , A Douglas | | Nicholas Gibbins |
| □ Resource Description Framework (RDF)<br>N Gibbins , N Shadbolt | 2009 | Nicholas Gibbins |
| □ User-Induced Hyperlinks in Collaborative Tagging Systems<br>CA Yeung , N Gibbins , N Shadbolt | 2009 | Nicholas Gibbins |
| □ Studies on Editing Patterns in Large-scale Wikis<br>P Boulain , N Shadbolt , N Gibbins | 2009 | Nicholas Gibbins |
| □ Cs aktivespace: Building a semantic web application<br>H. Glaser , H. Alani , L. Carr , S. Chapman , F. Ciravegna , A. Dingli , N. Gibbins , S. Harris , M. Schraefel , N. Shadbolt | 2004 | Nicholas Gibbins |
| □ Put in your postcode, out comes the data: A Case Study<br>T. Omitola , C.L. Koumenides , I.O. Popov , Y. Yang , M. Salvadores , M. Szomszor , T. Berners-Lee , N. Gibbins , W. Hall , N. Shadbolt | 2010 | Nicholas Gibbins |

⊞ Expand Found 1 publication by the name **N.M. Gibbins**

⊞ Expand Found 1 publication by the name **NM Gibbins.**

Claim Declaration

□ I have read the Terms & Conditions and I agree with it.

I declare that these publications are mine and I claim to be one of the authors of these publications and I am aware of any relevant copyright laws.

**Fig. 3.** User Auditing Interface. Caption (1)- Full Name Matching. Caption (2)- Exact Initial matching. Caption (3)(4)- Loose Initial Matching

*Full name matching* This makes two matches:

* It finds papers with an exact match of user's name with publication's author's name.
* It matches when author's first name starts with user's first name.

For example, author Nick Gibbins, Nick A. Gibbins can be matched with user profile name Nick Gibbins. We group these result together and pre-select them as it shown in Figure 3 point (1).

*Exact initial matching* Many authors' names in our dataset are not in full, instead, they are written in initial with their last name format. This finds papers that matches user's initial and last name. The initial of the user is computed by taking the first letter of the first name. We put these results in one group as it is illustrated in Figure 3 point (2) and it is not pre-selected because in most cases the results are from multiple authors. The figure also demonstrates a special case, where there is a user named Nicholas Gibbins who had already claimed some of the publications, the system highlights them to make distinction from the publications that are free to claim and the publications that have already been claimed by other users. If there is a wrongful claim or the claimed author is an impostor, user can follow the link to view the claimed author's profile details and can even report fraud.

*Loose initial matching* We take the initial and last name of all the authors in our database and match it with the current user's. This match finds authors that have multi-letter initials. As it is shown in Figure 3 point (3) and (4) there are two more matches - N.M Gibbins, NM Gibbins. We collapse this group of results for a cleaner interface, as there can be multiple results.

Some of our publication records also have email address associated with them, which can be a very accurate property to find user's publications. Therefore, we also enable users to enter all email addresses they use to publish their papers to do an automated pre selection of the paper as an option. For some special case, for instance when user has a different name associated with different publication, they can search the single publication and make a claim. This also holds true for misspelling or any other foreseen errors in the publications, user can simply search for them separately or add and even edit the publication themselves by the service 'Add or Edit Publication' provided by the system. When our publication database is updated or someone enters a new publication, users are notified to update their publication list as well.

In our system publications are modelled by using the Bibliographic Ontology [10] and the author of the publication is modelled by Dublin Core metadata[11]. The URI `http://www.iamresearcher.com/publication/rdf /1661006/` illustrates a single publication record. When user signup on our system, we generate a unique URI for each user and model their profile and their social relations by using FOAF ontology. When user claims a publication, they make alignment of the publication and their FOAF URI. These data is then pushed into the Linked Data Cloud and is dereferencable, e.g. by dereferencing the URI of this user `http://www.iamresearcher.com/profiles/id/yang2/` you will get an RDF file with list of the publications this user has claimed to be the author of. Following the publication links provided in the RDF an agent can easily pull out a user's co-author network graph and so on.

Our system is designed in a way that anyone can claim to be an author of any publication. Users are asked to agree to our terms and conditions as the

[10] `http://purl.org/ontology/bilbo/`

[11] `http://purl.org/dc/`

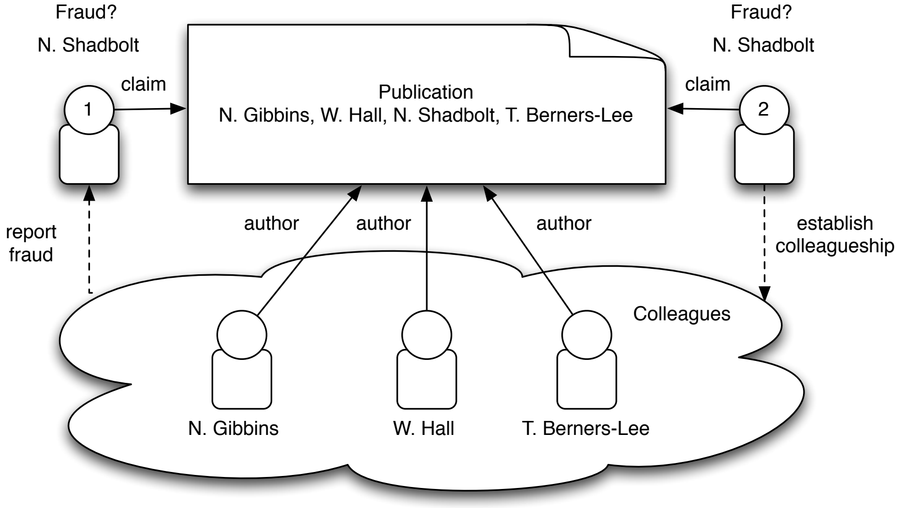


**Fig. 4.** Quality Cotnrol: Report of Fraud

system does not take any responsibility of breach of copyright or intellectual property issues, users who claim the publications are responsible for all the legal matters. So we enable users to report spam and fraud to maintain system integrity. The social network application has the benefit to identify a real user from fake by analysing their network structure and by examining the FOAF ontology. When dereferencing a user's FOAF URI, we get RDF to describe this identity, besides some basic information, we are defined by whom we related to. In our system, if a publication has five authors, it will be audited five times by all the authors. As we mentioned before, if you have co-authored a publication with someone, there is strong possibility that they are your colleagues too and you may want to establish the professional and social relationship as well. So in most of the scenario, we can easily identify a fraud because an impostor would fail to establish social relationships with other researchers.

Figure 4 illustrates how our system can spot a fraud. In this diagram, N. Gibbins, W. Hall and T. Berners-Lee claimed this particular publication, they have also established colleagueship in the FOAF file. Assume there are two users whose names are N. Shadbolt and they both claim to be author of same publication. How do we identify who is a fraud? As soon as one of them establishes a social relationship with any of the existing claimed author, others can spot and report the fraud. Indeed, someone can pretend to be someone to add social relationship as well, but it would be eventually spotted by observing day-to-day communication through the social network. For this purpose and for the convenience of users, we have designed a co-author invitation claim. So users can invite their co-authors to claim the publication and keep the integrity of the publication and in result the whole system.

## 6 Preliminary Evaluation Results

We deployed our system at the University of Southampton for evaluation. In this trial, we mainly focused on measuring two factors of the system: the incentive and total number of publications users claimed in comparison to the publication they deposited in the University Eprints repository. We sent emails to three different research labs at the University. We advertised the system as a free research platform and provided a link to the system. 163 users signed up initially (many of them were research students), we chose the 52 users who had deposited publication in the University repository as case study. As it is shown in Table 2, 39 out of these 52 users had claimed publications. This shows our system successfully used incentive as 75% users claimed their publications. Few research students gave feedback that they were pleased with the personalised recommendation system and easily found their publication. As our system aggregated data from the University repository (Eprints) and other repositories as well (1), our dataset is a superset of the University repository. By comparing the claimed publications, we estimated how well a user solved the co-reference problem in their own publications. In our analysis we found that, out of 39 users who had claimed publications, 51% of them claimed 136% more publications than they deposited at Eprints. It proves the success of the system as users claimed 39% more publications aggregated from different sources grouped together than the one where they entered bibliographical data in the Eprints repository themselves and also solved the co-reference problem in the integrated dataset. In contrast, 49% of users did not claim all of their publications they deposited in Eprints and only 67% of their publications were claimed. In the group of these users, we observed that many of them had a large amount of publications, for instance, one of the user had 417 publications, where he/she only claimed 289 pre-selected results. Claiming publications can be time consuming so our system also provides options for users to claim publication not only during registration process but also later at their own convenient time. Due to the time limitation,

| **Total Users(A)** | **Users have published(B)** | **Users claimed (C)** | **Percentage (D)** |
|---|---|---|---|
| 163 | 52 | 39 | 75% |

**Table 2.** User Claim Rate. Total Users(A): the amount of users registered to use our system within our University; Users have published(B): Amount of users who had publications in the University repository; Users claimed(C): From users in B, the amount of users who had claimed publications in our system; Percentage: C/B, percentage of users who had committed work to our system.

we did not observe the users for longer period to identify how many of them claimed publications later when system notified them to update their publication list. However, we believe if we deploy the system to a larger demographic, our system would produce even more promising results fuelled by the network effect.

| **No. of Users** | **Claimed Pubs(B)** | **Univ. Repos(C)** | **Claim Perc(B/C)** | **Perc of Users** |
|---|---|---|---|---|
| 20 | 1349 | 991 | 136% | 51% |
| 19 | 613 | 921 | 67% | 49% |

**Table 3.** Decomposition of users who claimed publications. Claimed Pubs(B): Claimed publications by that group of users; Univ. Repos(C): Amount of publications found in University repository for that group of users; Claim Perc(B/C): B/C, Claim percentage of that group of users. Perc of Users: Percentage of users who had made a claim. Since our dataset is a superset of the University repository, these users are presented with at least those publications that can be found in the University repository. If a user claims all the publication he deposited in the university repository, his Claim Perc would be 100%. This table splits those who under-claimed from the claimed authors. Therefore, those 19 users (bottom row), who on average claimed only 67% of the repository total, did not put in enough effort to find their publications in our system, while as the other 20 users (top row), who on average claimed 136% more that the university repository, managed to find publications we aggregated from other databases.

## 7 Conclusions

In some cases machines are not capable of solving the problems that are easier for humans and in our system we have taken the best of both worlds, the computational power of machines and cognitive ability of humans and brought them together to create a distributed human computation system to solve the Linked Data issue of co-referencing. This system creates an ecosystem by making the users, in this case researchers, the creators and consumers of data. Moreover, the platform we provide allows them to make a complete cycle of resource utilisation and consumption.

We have also emphasised the importance of incentive to motivate the user to contribute to the system and made a trustworthy structure to identify fraud and stop spam. But it is necessary to mention that since the system heavily relies on the human interaction and contribution, any shortcomings of humans is the shortcoming of this system as well. For example, if users fail to contribute then the system is unable to fix the errors, the system is as smart as the users using it and as diverse as the community of people. Also, as we take advantage of the researcher's social network, there is a risk of incomplete auditing of data when there are very few users from a network or research fields and we also need to take account of the old researchers who are no longer working and part of the research community or who are no longer active.

Finally, since the system is an ecosystem there must be equilibrium, the numbers of users and data are directly proportional, if there is scarcity of users or the data, the system will fail. And as it happens with any other natural ecosystem, it is vulnerable to unforeseen external factors and loose the balance to function properly as a stand-alone system. But with strong community contribution and support, the system can be resilient and thrive to become a stable and dynamic environment to provide better computation.

## 8 Acknowledgements

This work was supported by the EnAKTing project under grant EP/G0088493/1 and Web Science Doctoral Training Centre under grant EP/G036926/1, funded by the Engineering and Physical Sciences Research Council and the Research Innovation Services at the University of Southampton.

## References

1. J. Albors, J.C. Ramos, and J.L. Hervas. New learning network paradigms: Communities of objectives, crowdsourcing, wikis and open source. *International Journal of Information Management*, 28(3):194 – 202, 2008.
2. T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web: Scientific american. *Scientific American*, 284(5):34–43, 2001.
3. F. Feldman. Leibniz and" Leibniz'Law". *The Philosophical Review*, 79(4):510–522, 1970.
4. Hugh Glaser, Afraz Jaffri, and Ian Millard. Managing co-reference on the semantic web. In *WWW2009 Workshop: Linked Data on the Web (LDOW2009)*, April 2009.
5. Hugh Glaser, Tim Lewy, Ian Millard, and Ben Dowling. On coreference and the semantic web. December 2007.
6. Tom Gruber. Collective knowledge systems: Where the social web meets the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):4 – 13, 2008. Semantic Web and Web 2.0.
7. A. Jaffri, H. Glaser, and I. Millard. Uri identity management for semantic web data integration and linkage. In *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops*, pages 1125–1134. Springer, 2007.
8. In-Su Kang, Seung-Hoon Na, Seungwoo Lee, Hanmin Jung, Pyung Kim, Won-Kyung Sung, and Jong-Hyeok Lee. On co-authorship for author disambiguation. *Information Processing and Management*, 45(1):84 – 97, 2009.
9. I. Millard, H. Glaser, M. Salvadores, and N. Shadbolt. Consuming multiple linked data sources: Challenges and Experiences. November 2010.
10. Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 104–111, Morristown, NJ, USA, 2002. Association for Computational Linguistics.
11. A.J. Quinn and B.B. Bederson. A taxonomy of distributed human computation. *Human-Computer Interaction Lab Tech Report, University of Maryland*, 2009.

12. Manuel Salvadores, Gianluca Correndo, Benedicto Rodriguez-Castro, Nicholas Gibbins, John Darlington, and Nigel Shadbolt. Linksb2n: Automatic data integration for the semantic web. In *International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2009)*, June 2009.
13. N. Shadbolt, W. Hall, and T. Berners-Lee. The semantic web revisited. *Intelligent Systems, IEEE*, 21(3):96–101, 2006.
14. S.C. Shapiro. Encyclopedia of artificial intelligence, vols. 1 and 2. 1992.
15. Jennifer Sleeman and Tim Finin. Computing FOAF Co-reference Relations with Rules and Machine Learning. In *Proceedings of the Third International Workshop on Social Data on the Web*, November 2010.
16. Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.
17. Luis von Ahn, Manuel Blum, Nicholas Hopper, and John Langford. Captcha: Using hard ai problems for security. In Eli Biham, editor, *Advances in Cryptology 'EUROCRYPT 2003*, volume 2656 of *Lecture Notes in Computer Science*, pages 646–646. Springer Berlin / Heidelberg, 2003.
18. Yang Yang, Ching Man Au Yeung, Mark J. Weal, and Hugh Davis. The researcher social network: A social network based on metadata of scientific publications. In *Proceedings of WebSci'09: Society On-Line*, March 2009.