Writeslike.us: Linking people through OAI Metadata

Emma Tonkin

UKOLN, University of Bath, Bath, United Kingdom
e.tonkin@ukoln.ac.uk

Abstract

Informal scholarly communication is an important aspect of discourse both within research communities and in dissemination and reuse of data and findings. Various tools exist that are designed to facilitate informal communication between researchers, such as social networking software, including those dedicated specifically for academics. Others make use of existing information sources, in particular structured information such as social network data (e.g. FOAF) or bibliographic data, in order to identify links between individuals; co-authorship, membership of the same organisation, attendance at the same conferences, and so forth. Writeslike.us is a prototype designed to support the aim of establishing informal links between researchers. It makes use of data harvested from OAI repositories as an initial resource. This raises problems less evident in the use of more consistently structured data. The information extracted is filtered using a variety of processes to identify and benefit from systematic features in the data. Following this, the record is analysed for subject, author name, and full text link or source; this is spidered to extract full text, where available, to which is applied a formal metadata extraction package, extracting several relevant features ranging from document format to author email address/citations. The process is supported using data from Wikipedia. Once available, this information may be explored using both graph and matrix-based approaches; we present a method based on spreading activation energy, and a similar mechanism based on cosine similarity metrics. A number of prototype interfaces/data access methods are described, along with relevant use cases, in this paper.

Keywords: formal metadata extraction; social network analysis; spreading activation energy; OAI-PMH metadata; informal scholarly communication

1. Introduction

Social network analysis is frequently applied to study 'community' structures. Web 2.0, with its social nature, is expected to contribute changes to scholarly communication. Community data mining, i.e., mining real world-data such as scholarly articles in order to characterize community structures, is considered a top data mining research issue [1]. Automated inference through mining now represents a plausible approach to extracting candidate information from data that is already publicly and openly available in institutional repositories.

The Writeslike.us system was initially intended to support future work with the University of Minho in establishing informal links between researchers [2]. It also constitutes an exploration of a general area of interest, which is that of processing OAI metadata to extract author identity and to link it with external data (as in Wikipedia) and linked data.

Understanding the relationship between individuals and their communities is a very old problem. Several excellent tools exist that attempt to support the exploration of community structures within research publication information; a prominent example is the application of RKBExplorer [3] to support the exploration of OAI-PMH and DBLP metadata [4]. The problem has received widespread interest via bibliometrics and the potential for their application in the area of impact assessment for academic publications, as well as the widespread use of FOAF (Friend-of-a-Friend) to encode machine-readable information about individuals and their relationships.

One problem often encountered when making use of this sort of data is, in the case of FOAF, that the data is too sparse to provide an overview of the community as a whole; in the case of citation analysis, the information is much more exhaustive, but instead suffers from the difficulty that the data is not sufficiently extensive or accurate to enable a clean graph to be drawn. For example, FOAF information is necessarily limited to the data provided by individual users and their contacts on a given web site. Sometimes it is also possible to merge FOAF data from multiple web sites, but this is complicated by the need to identify equivalence between users.

FOAF information is generated by direct input from users; for example, creating an account on LiveJournal or Facebook creates a new individual identity. As the account holder seeks out individuals to add to their ‘friends’ list, they create new links between themselves and others. Characterising those links is sometimes difficult; some systems therefore ask users to specify the nature of the link. It is a simplification, based on the assumption that individuals play clearly definable roles in each others’ lives – but a social

graph can nonetheless be a useful resource. The problem, however, is that not all users make use of any given site, and a large percentage will make use of none at all. A recent report by Connell [5] offered a literature review of several studies demonstrating user opinion to social networking sites; one, from 2005, showed that over 50% of study participants responded that Facebook had no potential as an academic outreach tool, 12% said that it had potential, and the rest were unsure.

Reported demographics of use of social networks are largely consistent with the popular image of a bias towards younger users, although the picture is changing rapidly over time. A recent study [6] of the use of online social networks demonstrated that younger users were associated with higher levels of usage of Facebook, as well as a greater number of 'friends'. The types of use of the site also vary greatly between user groups, and therefore all social graphs are not created equally (nor equally detailed). Finally, the characteristics of different social networks vary in that some social networks predominantly reflect offline connections, whilst others are predominantly places where people meet for the first time online. Facebook, according to Ross et al. [7], is 'offline-to-online' – that is, Facebook friends are mostly met offline and then added at a later date to the online social network.

The outer edges of the Social Web

Citation analysis-based bibliometrics tends to privilege those papers that have a large number of citations or are published in journals or proceedings that are indexed in large citation databases such as the domain-specific PubMed, DBLP or ACM, or journal-specific but widely accessible citation indexes. Indeed, Mimmo & McCallum [8] point out that it is 'natural to ask which authors are most influential in a given topic'. Because citation is essentially a 'voting' process, giving those who are better-known or more influential or key to a given area of research a higher profile, less well-known authors and those who disseminate on a more local level will be almost invisible by those metrics, unless indexes also explore user-driven opportunities for deposit such as institutional repositories.

The majority of online bibliographic research tools, understandably, focus on establishing the primary figures within a field, and de-emphasise encouragement of collaboration between the 'foot-soldiers' of the research world. Yet there are reasons to explore this territory. The aim of this project is not to identify the most popular or well-cited individuals in a field, as there are many existing methods that enable this to be done. Rather, we aim to

explore a mixture of factors; matching expertise and interest and enabling a multifaceted browse model for information about individuals.

We chose to make use of a mechanism that depends only on data that is already publicly available, and as a consequence the startup cost was small. Additional data was extracted from publicly available sources, and is to be republished for others to reuse in the same spirit of encouraging innovation.

Service design

A number of well-understood components are required in order to create a system of this type. A data source and parser are required in order to extract the essential information – for example, author names, institutions, and other formal metadata. In the majority of cases citation analysis over a large corpus of electronic versions of papers is applied for this purpose, perhaps along with a formal metadata extraction system. We replace this step with extraction from OAI-PMH records.

Extracting strings from OAI-PMH records is extremely simple, but the difficulty lies with their interpretation. A perfect example of this is the author name disambiguation problem, which is to say, the question of identifying, from a pool of uses of a given string, which instances refer to a given individual.

The obvious conclusion when seeing five hundred papers by John Smith is that John Smith is a prolific author. However, in reality there are in all probability several John Smiths at work and writing peer-reviewed papers. The question becomes how to tell the works of a given individual from those of another individual with the same name. If there appears to be a Smith working in ethnography with a second working in quantum physics, then we are perhaps fairly safe in assuming that they are different people – but it is nonetheless possible that a single individual named Smith has moved from physics to HCI in the last few years. If one Smith works for Harvard and the other works for MIT then it is reasonable to assume that they are different people, but it is not by any means certain. Authors may be affiliated to several institutions; they may have moved from one institution to another, or taken a sabbatical to work in another and then returned to their parent institution. They may be working as a visiting fellow.

Author name disambiguation is a complex problem. Many unsupervised methods exist, such as calculating the distance between strings. We explored the use of several, including a vector-based cosine distance approach applying similarity between authors' coauthor lists, institutions and subjects (calculated using simple text analysis to extract approximate ‘noun phrases’, and then reducing these further using Wikipedia as a text corpus).

The most promising methods, according to Laender et al. [9], are generally based around supervised machine learning techniques, which require initial training. Examples of these include Naive Bayes [10], and support vector machines (SVM) [11].

On et al. also describe various heuristics to be applied across the names themselves; spelling-based heuristics, based on name spellings, token-based blocking, n-gram based blocking, and tokens similarity [11]. Laender et al. [9] describe a heuristic-based hierarchical clustering method that offers comparable results. It has been applied for a wide number of purposes; one that relates closely to our own is described by Minkov et al. [12], who applied contextual search and name disambiguation in order to relate name mentions in emails to a given identity.

Collecting relevant background information

During the development of the project identifying and utilising suitable data sources led to the need to make use of a whole variety of publicly available resources, such as Wikipedia; this required enhancement for effective re-utilisation of the data.

Many of the possible functions of this prototype depend on having access to appropriate datasets. For example, in order to improve the accuracy of a formal metadata extraction algorithm designed to identify the institution with which an author is affiliated, it is useful to have both a gazetteer of institution names and variant forms, and a list of the domains and sub-domains associated directly with that institution.

Again, this is greatly simplified when the data exists in a well-formatted, well-structured form, and to an extent this is true; for example, there exists partial data on DBPedia – a 'community effort to extract structured information from Wikipedia and to make this information available on the web' [13]. DBPedia is built up of the subset of information on Wikipedia that is well-structured and well-formatted, which is to say, predominantly information that is placed inside structured templates. However, although already a useful resource, there is not enough information to cover all of our requirements. In terms of institutions, for example, only a small subset of institutional Wikipedia pages contained a 'legible' dataset, and even in these cases, the information did not – by design – cover any of the variant forms of institutional name, identifier, domain, etc. that are useful for our purpose.

Therefore we chose to make a less subtle use of the Wikipedia resources, by spidering the pages directly, extracting relevant terms and URLs from the page source, and attempting to characterise them by means of application of a small set of heuristics. This was particularly difficult in shorter Wikipedia articles, articles that contained information about institutions that were not,

themselves, present on the Web, and articles that were written in languages other than English; however, we found that the simplest possible heuristic alone – that the top 1/3 of external links present on the Wikipedia page were likely to represent the institution – was correct over 65% of the time. We chose to take a permissive approach to information collecting and to add a confidence rating to each data point, reasoning that it is preferable to store too much data than too little at this early stage in the prototype's development.

Scenarios of use

Usage scenarios are real-world examples of how people can interact with the system being designed, and are often collected as part of development processes, especially user-centred processes. In general, usage scenarios are written with specific user personas in mind; in this case, they have been generalised for publication. Initial usage scenarios were developed through consultation with the University of Minho and institutional repository managers elsewhere. A shortened summary is included in this paper. Scenarios varied from supporting a student in seeking a supervisor working in an area of interest, to exploring influences of relevance to the modern-day concept of the impact factor. Several example scenarios that informed the design of the writeslike.us prototype are given here.

Scenario 1: Classifying events and forums by listed participants

A researcher in the field of evolutionary linguistics has become increasingly very interested in possible mathematical mechanisms for describing the nature, growth and adaption of language, as he has heard that others have done some very interesting and apparently relevant work in this area. Unfortunately, the researcher finds that some of the detail is hard to follow. He decides to seek out an appropriate event and/or online forum, and finds some people who might be interested in exploring links between his specialist area and their own. He is concerned about the potential cost of attending several events, so he chooses to look up possible events and web forums, intending to look through the participant lists for names that he recognises. This is greatly simplified by an automated system enabling him to identify papers and authors that he considers most relevant; with this information it is possible to parse through lists of participants in events or online communities in order to provide him with a rough classification of how relevant the group is likely to be to his ideas.

Scenario 2: Building a 'dance card' for a Research Council event

One of the purposes of a Research Council event is to encourage serendipitous meetings. Rather than simply assuming that synchronicity at the coffee-table will carry the day, the Research Council decides to produce a 'dance card' that suggests several other individuals that you might like to meet. Whilst elements of the composition of this 'dance card' are resultant from program managers' knowledge of the individual's interests and character, a service that is able to identify, characterise and compare researchers from their existing work and affiliations can be used to quickly build some interesting (and at times amusing) meeting suggestions, based on the individuals' papers and output, and/or on the names and Research Council-held descriptions of the projects on which the individuals work.

Scenario 3: Facilitating collaboration in a multidisciplinary research environment

An anthropologist with a particular interest in the area of paleolithic archaeology, who works in the Department of Humanities, is very interested in exploring likely patterns of migration, and particularly in the idea that this activity may have been driven by climate change. However, the Department of Humanities has limited funding for the purpose of data collection and interpretation regarding modeling of climate change, so it is not possible for him to develop a paleoclimate simulation system. Therefore he decides that it is more appropriate for him to look for other people who have other reasons to be interested in modeling of this kind, particularly during the time period in which he is interested. This is not a trivial problem for several reasons; firstly, he does not usually publish in the same area as paleoclimatologists and therefore is unlikely to make chance acquaintances. Secondly, he and they have very different ways of describing their areas of interest, and therefore there is quite a lot of interpretation required in order to ascertain that the datasets they require are (or are not) closely related. Successfully establishing that these groups could usefully share data with each other is a non-trivial problem. However, it is an important goal for all concerned, not only because it is likely to help the data *consumer* – the anthropologist – but also because the data *creator* – the palaeoclimatologist – will benefit from wider impact and reach of their research.

Engineering from a series of scenarios

It is noticeable that the majority of the scenarios depend on information taken from several sources; much of the benefit of this system requires the data extracted to be enriched with externally sourced information.

For example, geographical information is necessary in order to successfully complete some of these tasks, such as finding local academics with similar interests. In principle, this information can be found from several sources, such as the address of the institution that the author places on his/her conference submissions. However, affiliation with an institution does not necessarily require that the author spends a great deal of time physically present at that institution. In practice, reliably eliciting a researcher's workplace or present location is difficult, requires considerable information to be made available – such as calendar data, GPS, etc – and is not really practical without finding a solution for several major infrastructural and social issues.

This additionally marks an intersection with well-known research themes in the field of ubiquitous and pervasive computing, in particular the many research projects that have sought to enhance social networking by means of additional information gleaned from context-sensitive mobile computing. For example, Kortuem and Segall [14] describe a system that supports augmentation of social networking through mobile computing. Information such as the individuals with whom people have met and spoken can be collected and stored [14]; individuals can 'pledge' to work together towards a given aim, and discovery of other individuals in the group can be managed accordingly. This form of mobile computing requires devices that are always on and running (*constant*), aware of presence of nearby devices and people, able to communicate with other collocated devices (presumably a heterogeneous environment of devices), and proactive – operating without explicit user interaction. All this goes to underline the point that, although there is great potential in this sort of work, it is non-trivial as an engineering problem.

Because so much data is involved – or at any rate, sought – there is great potential for this sort of work to be implemented via the reuse of data from several other sources, in what is sometimes known as a mash-up. That said, much of the data is not sufficiently cleanly formatted or complete to be accessible as simply as well-formatted and structured FOAF, meaning that the major challenge here is one of extracting as much as possible from the information available and storing it in a normalised form. The second challenge is to take the data and attempt to establish equivalence between

entities, which is to say, explore whether multiple mentions of the same name refer to the same individual.

As a result, this problem can be seen as a useful step in the general problem of author name/identity disambiguation. Much of the data mined could potentially be used to enrich formal data sources such as that offered by the NAMES project [15].

2. Methodology

The writeslike.us project contained several specific stages; harvesting, analysis, user-level interface development and evaluation. . The project had as its final goal the development of a functional prototype intended to extract as much information as possible, making it accessible for further research in the area.

Source data was retrieved from OAI-PMH metadata repositories, but was expected to require supplemental information from several sources; identification of appropriate sources was itself an aim of this work. As such, one focus of the project was to identify and use existing services wherever possible. The quantity of data was also expected to give rise to a number of data management issues.

2.1 Data collection

The dataset is harvested from OAI-PMH – the Open Archives Initiative Protocol for Metadata Harvesting [16]. According to the Repositories Support Project [17] about 75% of institutional repositories worldwide, and ~85% in the UK, provide an OAI-PMH interface. The dataset is harvested via UKOLN's OAI data harvesting service, RepUK, which performs regular UK-wide data harvests and stores the resulting information in an XML dump, available for reuse by other services and applications. The system currently takes in data from across the UK - a future evaluation methodology will be to explore the applicability of the system at an international level. The data is initially input into a database, from which it is processed in the following manner:

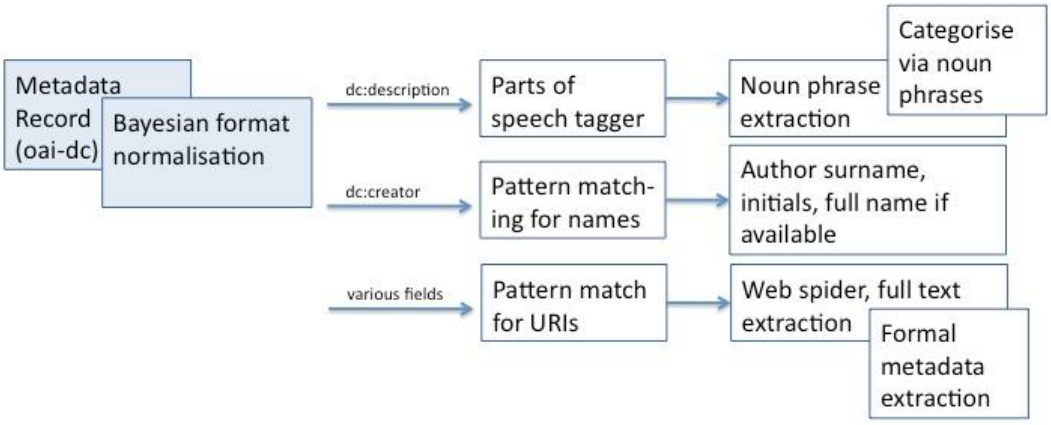

Figure 1: Published articles per year

2.2 Collecting information from the metadata record

The metadata record itself is collected in oai-dc, which is to say that it contains up to twelve elements, most of which are unlikely to be filled out. The actual placement of information within the record may vary a great deal depending on the interface used to create the record and the design decisions of its originators. Often the information will have initially been input as qualified Dublin Core rather than the simpler format, and as such there are likely to be several instances of certain fields, containing different refinements of a given field, but not marked as such. Hence, a useful initial step in analysing this information is to attempt to characterise its qualified representation.

This step can be achieved in a number of ways; for example, a simple heuristic can often identify certain fields such as well-formatted dates (ie. validation against a schema, regular expressions capable of identifying common citation formats, etc). Content-level feature analysis can also help to identify certain common types. The result, however, will not be as clean as collecting the qualified information directly in the occasions on which it is available.

2.3 Format normalisation

Knowing what type of information is contained in a given field is only the beginning. Following that, it becomes necessary to come to an understanding of the format of that information; for example, author names may be given in any one or more of a number of different formats:

Smith, John and Richard, Peter
J. Smith and P. Richard
John Smith, Peter Richard.
Smith, John and Peter Richard
SMITH, J. ; RICHARD, P.
…and so on.

It is unlikely that all of these can be convincingly, uniformly and accurately parsed, especially since there are certain cases in which there is an essential ambiguity that would stop even a human user from having any certainty of his or her conclusions – for example, Peter Richard and Richard Peter are both acceptable, valid names. A Bayes filter designed to make use of a knowledge base of information regarding the prevalence of different first names and surnames would have a similar problem in this case; making use of the last published set of US census data, a classifier might make use of Table 1 to find the following statistics:

Table 1: US Census data for surnames Richard and Peter

Name	Rank in US	Approx. number in US	Freq. of occurrence per 100,000
PETER	3758	8662	3.21
RICHARD	581	52138	19.33

True, 'Peter Richard' is a more statistically probable match than 'Richard Peter', but neither solution has an overwhelmingly convincing lead over the other.

Very similar problems exist with many other data formatting areas, for example, dates; the well-known discrepancy between preferred forms of date formatting in the United States versus the preferred mechanisms in use in the United Kingdom mean that:

10-11-2009
11-10-2009

may be very difficult to parse. Ideally, these formatting and encoding problems would be avoided by appropriate choice of standard and strict adherence to that decision. As these issues do occur in practice, we are left with the requirement of developing a mechanism for solving such problems with reasonable degrees of success. Fortunately, there are often simple heuristics that can be applied. In both of the cases described here, there is a

solution to do with the observation that these variations are often regional and often have a link to the choice of interface in use. If this is the case, then the convention is likely to be at least somewhat systematic. It then becomes desirable to identify the convention in use within that context, the frequency with which it applies and the likelihood that a similar pattern holds in our particular case.

2.4 Categorisation via text analysis

The means of categorisation was described briefly above as a method of extracting noun phrases from text, followed by dimensional reduction using Wikipedia as a corpus. Here, we will explore how this works in more detail.

Noun phrases are defined by Wordnet as 'a phrase that can function as the subject or object of a verb'. We do not precisely look for 'noun phrases', but content ourselves by looking for a more loosely defined phrase that is either a noun phrase or a descriptive phrase of a somewhat similar nature. This is achieved by making use of a part of speech tagger to analyse the textual content that is available to us. This is naturally language-dependent, and marks the point at which it is no longer possible to work in a language-agnostic manner. The particular tagger that we used is available in a number of common European languages (English, French, German, Spanish, Portuguese).

This is a relatively processor-heavy and slow part of the analysis process. To complete a single run of data analysis on the full textual content of the metadata records alone takes over 24 hours – this is on UK content only, which is to say, a few hundred thousand records. To perform a similar task on a global scale would take weeks. To run a similar analysis on the full-text of each document spidered (see the following subsection) would take much longer. As such, it is important to consider both efficiency and the possibility of caching results – if a piece of data has been processed once, that information should be stored for the future.

In practice, the data set proved to be large enough to feed quite an extensive collection of terms, meaning that there was enough information to use this information for browsing purposes. Smaller datasets offer difficulties both to interface designers and for those intending to reuse the information for categorisation purposes, not least because there is not sufficient information regarding similarities between documents to enable an effective categorisation process to occur.

However, the data did suffer from excessive specificity; that is, there are similarities between the concepts underlying 'superstring theory' and 'Higgs boson', but it is not necessarily obvious from the content of the term or of

their application. This problem can be tackled in a number of ways. One common solution is to make use of the LSI approach (latent semantic indexing) to reduce the dimensionality of the large matrix of terms, in the hope that the similarities between term domains would become evident. However, we chose to make use of Wikipedia as a 'crowdsourced' text corpus on which to look up and attempt to identify classifications for each of the terms that we had extracted. This meant that terms related to, for example, physics, would be clearly identified as such.

Once the popular phrases are extracted, they can also be used as navigational elements supporting a subject-level browse mechanism. The same is true of dimensionally reduced categories, although in this case it may be preferable to treat them as a simpler set of categories, graphically presented in a similar manner to a standard 'breadcrumb trail' navigational element.

2.5 Formal metadata extraction

Before formal metadata can be considered, it is necessary to identify appropriate candidate data objects from which that metadata can be taken. For example, an OAI record may refer to a paper, a presentation (.ppt, etc) file or a dataset contained within the institutional repository. These digital objects may not be accessible to the outside world – indeed it is not uncommon for records to be placed online without depositing a data object. Alternatively, the object(s) may be hidden for a limited time (embargoed).

Because direct links to digital objects are rarely given within an OAI record (only ~600 records contained actionable, externally accessible links directly to digital objects), it is often necessary to use a web crawler (spider) to identify candidate links to the fulltext record. About a quarter of the records surveyed contained a link to a page from which the originating record could be retrieved. To harvest digital objects effectively can be time-consuming and the details are outside the scope of this paper, so we will merely comment here that whilst there is considerable variation to be handled, handling the most common is easily achieved.

We found that on average just under half of the pages retrieved through crawling of links provided within DC records contained one or more accessible documents (Fig. 1). Around 15% of linked pages resolved to a variety of journal endpoints – 'paywalls' - (Ingenta, Taylor & Francis, Wiley, Sage, IOP, etc). These sometimes contain additional useful metadata about the document, but do not contain the document themselves. However, the copyright ownership is in itself a useful data point. Around 40% of institutional repository links were found to contain no accessible data.

A further finding from this work was the number of DOIs and handle.net persistent links that were present in metadata records, including the percentage that were broken. 240,000 records were harvested. Out of the 62,000 records containing an actionable http dc:identifier, 35,000 contained a handle.net (15,500) or dx.doi.org (20,000) actionable persistent identifier. DOIs and handles appear to have a similar prevalence in UK institutional repositories.

In principle, persistent actionable identifiers are useful in part because they permit URLs to be assigned, managed, and reassigned when something causes that persistent identifier to break. In practice, we found that there was a noticeable proportion of broken persistent identifiers. We found that out of a test run of 20,000 URLs retrieved from URL records, almost 400 were unresolvable DOIs or handle.net persistent identifiers – 2% of URLs were invalid persistent identifiers, of which two-thirds were DOIs.

Overall, harvested data is retrievable for about 1/8 of the indexed objects. Additional metadata (for example, details about a journal publication, author, affiliation, citations, and so on, that are not reflected in the metadata) may potentially also be retrievable from journal item pages, so there is a case to be made for this approach.

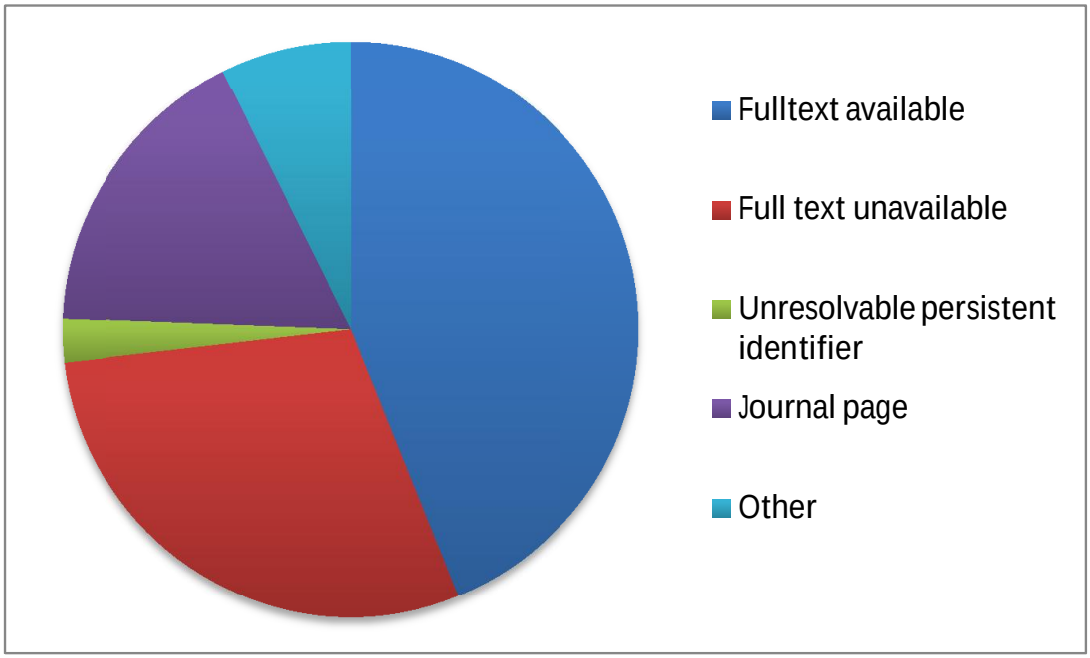

Figure 2: Findings from web crawl

The sorts of information that can usually be retrieved from this approach include, for eprints: title, author name, sometimes a date or range of dates, citations, format information, format metadata, software from which the

document was created, and perhaps additional information such as keywords, abstracts, etc. This provides a useful check against the OAI metadata to ensure that the correct document has been retrieved. For most digital objects it is possible to retrieve some form of format information and/or object-specific metadata. Object provenance information such as software used in its creation can say a great deal about author identity, as specific authors may choose different routes to object creation.

2.6 Name disambiguation in metadata graphs

The result of the data collection process is to develop a set of relations between entities, author-name strings and articles. Before the information can be effectively used, the distinction between author name strings and identity must be applied. Characteristics of authors for a given paper may often be extracted directly from the article text. These pieces of information may be added to what we already know about a given set of instances of use of a string. Several name disambiguation approaches were explored using test datasets taken from the real-world OAI-PMH dataset.

The information that is available to us can be displayed as a matrix of terms; a given author-string *a* in authorlist *A* has created i_a digital objects out of an overall set of objects *I*. Each object has a set of subject characteristics s_i, a title t_i, author affiliation information f_i, provenance from repository r_i, and so forth. So the problem becomes one of ascertaining the most likely number of individual authors identified by string *a*. These characteristics can be treated either as links within a graph, or a sparse matrix of characteristics; both graph-based approaches such as spreading activation energy and matrix-based methods such as cosine similarity metrics may be applied.

Table 2: Document object data in matrix form

I	*t*	*f*	*r*	*s (a, b, c…)*
i_2	t_2	f_2	r_2	s_2*(a,b,c…)*
i_n	t_n	f_n	r_n	s_n*(a,b,c.…)*

2.6.1 Relatedness metrics: Cosine similarity

The first metric explored was a very simple cosine similarity calculation. This is a measure returning the similarity between vectors of *n* dimensions, and is a staple of search engine design [18]. Cosine similarity states that two identical vectors are exactly identical (eg. Eq. 1 returns 1) or that they are dissimilar (Eq. 1 returns 0).

$$similarity = \cos(\vartheta) = \frac{\mathbf{A}.\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

Equation 1: Cosine similarity

Cosine similarity can be applied to the document object characteristics almost directly, although there is a need to 'unpack' listed characteristics into a simpler numeric form if they are to be treated as term frequencies. Applying this directly to information such as that provided in Table 2 returns a modified form of document object similarity, weighted with additional factors such as provenance metadata. To relate this to the author requires an assumption that there is a relationship between document object similarity and author identity. In principle this would seem to be a strong assumption, - authors primarily write about topics that they know well and seldom write outside their field. However, in practice this assumption may not hold up over time, as authors may change research groups, areas of interest, and even subject areas. It may also presuppose a meaning to authorship that is not there; for example, co-authorship of a paper may imply that the author has produced data that was used within the paper, but the paper itself may not be in the author's core area of interest – a palaeoclimatologist may co-author a paper with an archaeologist, published in a journal with a focus on archaeology and written for an audience consistent with the journal's focus. Authorship represents contribution, which is traditionally expected to be textual, but often this may be an experimental collaboration, inspiration, supervision, perhaps even an administrative link.

This method provides a reasonable metric for comparison between items. There is, however, considerable computational overhead in calculating this for each document and metadata set, so this was done periodically in exploring the link between author name and unique identity - that is, this method provided one data point for the following question, essentially a clustering problem without the advantage of any definitive knowledge of the number of authors involved: given a set of objects created by a person or people with the author name *a*, how likely is it that objects *i* and *j* were created by the same author? Note, however, that grounding author identity as a subset of author name ignores a large number of complications, such as authors who change their names or anglicise their names inconsistently.

The exploration of similarity provides an approximate notion of identity, which can also be represented as nodes in the graph – which, indeed,

supersede author name string similarity in calculating the relatedness between papers.

2.6.2 Relatedness metrics: Spreading activation energy

A search algorithm based on spreading activation energy over a contextual network graph modelled over a series of timesteps [19] [20] was applied to support graph-based searching. This is not a pre-calculated method, but is rerun on each search. This method is appropriate where it is possible to search from the starting point of a prechosen node – given a unique node id as a search key, this algorithm could be seen as spilling a little ink on one node of the graph, which then spreads a predefined distance through the graph of relations between authors, objects, roughly calculated identities, classifications, and other metadata, in a manner defined by the way in which the implementation is tuned. The result is a ranked list of matching nodes and their types, which can then be presented to the user.

Modification of this approach can be used to reflect the relative relevance of different types of connection or to tune a search to prioritise different types of relation (eg. topicality, similar location of publication, similar physical location). The search may be weighted according to specific search types. In terms of efficiency, frequent calculation via a contextual network graph algorithm is observed to be relatively inefficient on a dense graph, by comparison to alternative methods (eg. latent semantic indexing); intuitively, this is reasonable, the number of links to process per timestep is related to the density of the graph. The decision of whether to use a contextual network graph/spreading activation energy method or whether to precalculate is also linked to the number of changes expected to be made to the graph – frequent change and hence frequent recalculating negates any benefit to be gained from what is essentially a caching mechanism. Furthermore, the contextual network graph approach is memory-hungry [19]; in a production environment it may be preferable to pre-process the data in a persistent (cached) form.

3. Evaluation

Initial qualitative and quantitative evaluation studies have been completed, and preliminary results are encouraging. In particular, there is significant diversity between information and authors indexed into well-structured datasets such as DBLP, ACM, and so forth, and the world as viewed through

OAI-PMH. From a random sample of authors, it is very visible that authors with few publications have little visibility in the formal indexes. Figure 2 contrasts authors who have published between six and twenty papers with the indexing visibility of authors who have deposited five or fewer documents (see Figure 2).

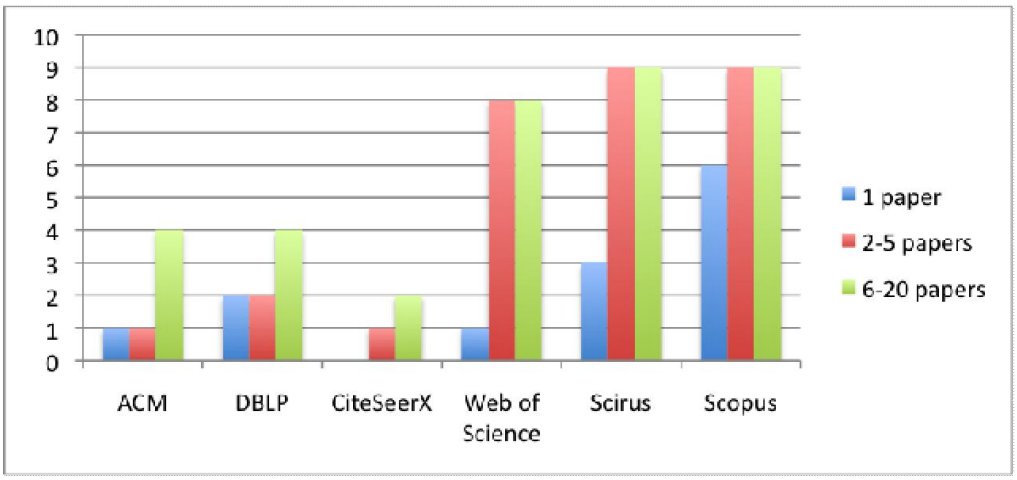

Figure 2: Author appearance in popular indexes (average values)

One potential interpretation of this limited visibility is the following: it could be taken to indicate that the data is relatively low-quality, as it contains so much information that was, for whatever reason, not published in a popularly indexed publication – and hence perhaps it is not published in a peer-reviewed form. However, for the purpose of encouraging informal collaboration between researchers, this may not prove to be a significant impediment.

Certain areas of functionality are key, in particular those areas concerning author identity, and it is clear that alternative methodologies, the use of additional data sources, and user-level amendment functionality could improve things greatly.

4. Discussion

During this work, we established a database extracted primarily from metadata, a knowledge base derived from Wikipedia, a variety of classification information derived variously from part-of-speech tagging and the use of Wikipedia as a classification dictionary. We used this in order to retrieve further information where available, and extract what metadata was available from these sources. We then used simple mechanisms from text analysis to classify and provide a mechanism for exploring this data.

The major difficulty in the project was simply one of interface design and access; it is one thing to develop a database, and quite another to create an interface that supports the aim of encouraging informal collaboration – an aim that depends a great deal on factors that are very difficult to identify or represent digitally, such as trust and organisational culture [21]. The problem is as much social as technical. We expect to release parts of the information extracted as linked data for future reuse by ourselves and by others. One likely future avenue for reuse may be the application of this data in supplementing more formally generated information sets – 'filling in the gaps'.

The course of this experimental development has been a journey of discovery, and not least an opportunity to challenge our own assumptions about appropriate interpretation of apparently straightforward data, even those as simple as document creation and authorship.

5. Conclusion

OAI-PMH metadata alone provides sufficient information to collect basic information about authorship data. However, the quality and completeness of that data is greatly improved if the full-text document is also available and may be analysed. Another means of supplementing basic data about authors is to compare and contrast with information derived from networks of citations; for popularly-cited texts, this permits the retrieval of additional information such as authors' full names, relevant dates and so forth. The effectiveness of this approach is dependent on the number, style and quality of existing citations - so again, a synthesis of available approaches is likely to provide the best overall result. Methods of disambiguating unique author identity vary greatly in effectiveness depending on the available data, as well as factors such as resource type, format and language. Standardised benchmarking requires the establishment of and testing against a ground truth. Bibliographic networks are often used to identify 'star' researchers working in each field. The problem explored here is to enable the ability to browse for others (who may be at any of various stages in a research career) working on a given topic area.

In our future work, we intend to widen the availability of writeslike.us as a pilot service, to develop a clearer set of requirements for practical usage of the interface, API and data, and to explore questions such as automated classification of authors' likely primary occupations (eg. primary investigator, researcher, technical writer, student). We also intend to explore the

possibility of bringing in other sources of data – and publish relevant segments of existing information for wider reuse, as a clearly and consistently formatted linked-data resource.

Acknowledgements

The author would like to acknowledge the input of Alexey Strelnikov, the primary developer of the web interface and database, Andrew Hewson, database guru, and Wei Jiang, for his work in the area of identifying and testing possible author name disambiguation methods.

Notes and References

[1] VARDE, A. Challenging Research Issues in Data Mining, Databases and Information Retrieval. In ACM SIGKDD Explorations Journal, 11(1) 2009, p. 49 - 52.

[2] BAPTISTA, A; FERREIRA, M. Tea for Two - Bringing Informal Communication to Repositories. D- Lib 13 (5/6) (2007). Retrieved 10th Jan, 2010, from www.dlib.org/dlib/may07/baptista/05baptista.html

[3] GLASER H; MILLARD I; JAFFRI A. RKBExplorer.com:A Knowledge Driven Infrastructure for Linked Data Providers. In: European Semantic Web Conference, 1-5 June 2008, Tenerife, Spain. pp. 797-801.

[4] GLASER H; MILLARD I; CARR L. RKBExplorer: Repositories, Linked Data and Research Support. In: Eprints User Group, Open Repositories 2009, 20/05/2009, Atlanta, GA, USA.

[5] CONNELL RS. Academic Libraries, Facebook and MySpace, and Student Outreach: A Survey of Student Opinion. Portal: Libraries and the Academy. Volume 9(1), January 2009. E-ISSN: 1530-7131

[6] JOINSON AN. 'Looking at', 'Looking up' or 'Keeping up with' People? Motives and Uses of Facebook CHI 2008, April 5–10, 2008, Florence, Italy.

[7] ROSS, C; ORR, ES; SISIC, M; ARSENEAULT, JM; SIMMERING, MG; ORR, RR. Personality and Motivations associated with Facebook use. Computers in Human Behaviour. Volume 25 (2), March 2009. Pp. 578-586.

[8] MIMNO, D; MCCALLUM A.s Mining a Digital Library for Influential Authors. JCDL 2007.

[9] LAENDER AHF; GONCALVES MA; COTA RG; FERREIRA AA; SANTOS RLT; SILVA AJC. Keeping a Digital Library Clean: New Solutions to Old Problems. DocEng '08.

[10] HAN H; GILES L; ZHA H; LI C; TSIOUTSIOULIKLIS K. Two Supervised Learning Approaches for Name Disambiguation in Author Citations. JCDL 2004.

[11] ON, B-W; LEE, D; KANG, J; MITRA, P. Comparative Study of Name Disambiguation Problem using a Scalable Blocking-Based Framework. JCDL 2005.

[12] MINKOV E; COHEN WW; NG AY. Contextual Search and Name Disambiguation in Email Using Graphs. SigIR '06, Seattle, Washington, USA.

[13] JENTZSCH, A; et al. About DBPedia. Retrieved March 20, from http://dbpedia.org/About

[14] KORTUEM G; SEGALL Z. Wearable Communities: Augmenting Social Networks with Wearable Computers. IEEE Pervasive Computing, volume 2(1) 2003, p. 71-81, ISSN: 1536-1268

[15] NAMES project introduction. Retrieved March 18 from http://names.mimas.ac.uk

[16] LAGOZE C; VANDESOMPEL H. "The Open Archives Initiative: Building a Low-Barrier Interoperability Framework". Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01). pp. 54–62.

[17] Introduction to the RSP project. Retrieved March 18 from http://www.rsp.ac.uk/usage/harvesters

[18] TAN, P.-N; STEINBACH, M; KUMAR, V. *Introduction to Data Mining*. Addison-Wesley (2005), ISBN 0-321-32136-7, chapter 8; page 500

[19] CEGLOWSKI M; COBURN A; CUADRADO J. Semantic search of unstructured data using contextual network graphs. 2003.

[20] BOLLEN, J; VANDESOMPEL, H; ROCHA LM. Mining associative relations from website logs and their application to context-dependent retrieval using spreading activation. 1999.

[21] DODGSON, M. Organizational Learning: A Review of some Literatures. Organization Studies, 14(3), 375-394 (1993). DOI: 10.1177/017084069301400303