Methods for Mining Messy Real World Data: Co-reference Identification Using Fuzzy Logic

by Stephen Brown, David Croft and Simon Coupland

Citation


Abstract

As the number and volume of online museum collections grows, there is an increasing imperative to improve their discoverability by finding ways of linking records that go beyond simple keyword searching. Keyword searches are inefficient because they are prone to errors of both omission and commission. A range of more sophisticated approaches have been developed for cross collection searching including metadata harvesting, data mining, Linked Data and Application Programming Interfaces (APIs) but these variously rely on availability of well structured, consistent and standardised data and a large corpus of text. While some heritage institutions have successfully implemented one or more of these approaches, pioneering the way for others to follow, the majority of online collection records are not amenable to such treatments because they employ different data schemas, which when they are applied inconsistently, are often fragmentary and imprecise and non-machine readable. A further challenge is that while there are many millions of individual object records, each containing many different fields, the entries in each separate field tend to be quite small, eg. dates, person names, titles, making it difficult to apply corpus based approaches to data analysis. Converting fragmentary, messy museum records into well-structured data is an unlikely prospect in the immediate
future because of the costs and technical skills required. This paper describes an alternative approach to handling messy real-world data, first introduced at the Digital Humanities Congress in 2012\(^1\), and reports on the results. We are using a constrained hierarchical clustering approach on top of a fuzzy inferencing system that can accommodate messy real world data. Record pair similarity is calculated using four separate similarity metrics, with each metric tuned to the specific challenges of the information held in that field, and an overall record similarity metric to combine the other four. Co-reference matches have been identified for over half the target records, enabling us to see some of the photographs shown at the exhibitions of the Royal Photographic Society for the first time in over 120 years as well as many more similar photographs.
Methods for Mining Messy Real World Data: Co-reference Identification Using Fuzzy Logic

by Stephen Brown, David Croft and Simon Coupland

1. The Challenges of Museum Data

In recent years there has been rapid growth in the quantity of digital resources placed on line by Galleries, Libraries, Archives and Museums (GLAMs). As of January 2015, the Victoria and Albert museum (V&A) website\(^1\) offered access to over 1.1 million works in its collection and the British Museum offered access to more than 2 million records online.\(^2\) These are not isolated occurrences. In 2006, a survey of 18,142 museums and libraries by the Institute of Museums and Library Services in the USA found that the majority of museums and larger public libraries, as well as a smaller proportion of public and smaller academic libraries already made digital records available to the public via the Internet (IMLS, 2006). Collectively these resources amount to a vast quantity of data. The Europeana Aggregator website alone offers access to around 30 million objects from over 2000 GLAMs.\(^3\)

The availability of such enormous quantities of digitized content creates both opportunities and problems for researchers. On the one hand it creates tremendous potential for making connections between records to construct new understandings and for using information about objects from one data set to enrich records in another, thus enhancing the value of both. On the other hand the sheer volume of information is daunting, placing the goal of comprehensive, global, cross-collection searching beyond the bounds of human capability. Of course computers can help. The Internet is served by powerful search engines and other tools that help users to process millions of records to find, filter, aggregate and share online data, but there are limits to the capabilities of routine search engines. GLAMs tend to hold their records in a variety of different content management systems, structured, labelled and marked up in various and sometimes idiosyncratic ways that reflect differences in the materials themselves, the purpose of their host institution, local needs and individual preferences and skill levels among
those doing the cataloguing (Dearnley, 2010; Eklund, et al., 2010; Henry & Brown, 2012; Kamura et al., 2011). This is not usually a major problem for researchers sifting through different collections on their own, but lack of consistency is problematic for computer programs that need to be told what sort of data they are processing, how to read it and how it is structured and labelled.

As if this were not problematic enough, GLAM’s object data is often encoded as natural-language text which is not readily machine-readable.

Consequently, simple keyword searches using ordinary search engines are insufficient for discovering important relationships within large quantities of museum records sourced from different institutions. Such searches typically return either impossibly large numbers of results that include a high proportion of irrelevant content –so-called errors of commission–, or very few hits that overlook important items –errors of omission–. Moreover, much of the Web lies beyond the reach of search engines because it resides in databases and is presented dynamically to the Web only in response to specific queries. Museum object records are usually published this way, as evidenced by a recent survey of 140 digitised collections (Stephens, 2013) that reported that collection level data is much more discoverable than individual items from the collections. Whilst almost 100 per cent of collections were in the top ten Google results using the collection or project name as a search term, only about 50 per cent of items appeared on the first page of Google results using the item name or title. It seems therefore that we need to develop more effective tools “to navigate among vast catalogues of born-digital and digitised materials, as well as the records of physical materials” (ACLS, 2006). Some GLAMs have begun to explore the use of Application Programming Interfaces (APIs) for opening up collections to cross-collection resource discovery (Dearnley, 2011; Morgan, 2009; Ottevanger, 2008; Ridge, 2010). APIs are pieces of code that specify how programmes interact, along the lines of “if you go here, you will get that information, presented like this, and you can do that with it.” (Ridge, 2010). APIs are useful for enabling data mashups that combine object records with geographical data to create interactive maps for example, or for searching within database-driven Web sites. Many major Web data sources such as Google, eBay and Amazon provide access to their data through Web APIs. APIs are of interest to us here because they can be used to enable institutions to share data from different sources. Their principal advantages are that APIs provide access to source data without having to copy or move the data itself. That means there is no need to keep a separate copy of the
data in synch with the original, and material, once created, can be published in a variety of forms and locations with little further effort. It also means that APIs are fast, delivering results “on the fly”. Some examples of successful API implementation are the Science Museum, London (Ridge, 2010), the V&A (Morgan, 2009), the Powerhouse museum and Brooklyn museum (Dearnley, 2011) and the Rijksmuseum (Jongma, 2012), although these support in-house developments such as the V&A “Search the Collections” service and interactive exhibits throughout the galleries rather than services across multiple institutions. However, there are memory institution aggregator services, notably the Culture Grid and Europeana, that can ingest records via APIs. In the UK, the Culture Grid is effectively an API for cross-institutional collaborations involving circa 3 million records from hundreds of different collections and Europeana is a still larger scale portal, with access to 30 million object records.

The requirements for APIs that will work across multiple collections constitute demanding challenges for many GLAMs:

- Records must be consistently structured, unambiguous, clearly labelled and encoded in a machine readable format, such as XML.
- The institution must have the technical capability to write the code for (or buy in) APIs that will expose their own data.
- APIs must be clearly documented such that another developer can understand what the API does and what encoding requirements it has in order for another programme to interface with it effectively.
- Precautions need to be taken to protect institutional data and the institutional Web site.
- The access mechanisms and the format of the data retrieved from different collections needs to be standardised.

Another potentially promising approach is the concept of Linked Data (Bizer et al., 2009; Heath and Bizer, 2011; Oommen and Baltussen, 2012). The W3C consortium defines Linked Data as “a Web of Data (as opposed to a sheer collection of datasets)…” that works by defining the meaning of words that appear inside Web documents, and by explicitly linking those words to external data sets that help to define their meaning, so as to create a web of interconnected terms (Bizer et al., 2009). Linked Data thus underpins the functionality of the Semantic Web which “…. is about making links, so that a person or machine can explore the web of data.” (Berners-Lee, 2006). Machine-readability is key to the success of Linked data and there are Linked Data search engines that can query data inside Web documents eg.
Falcons, SWSE (Bizer et al., 2009), provided it is expressed correctly.

Linked Data is encoded in the form of so-called ‘RDF triples’ that comprise a subject, predicate and object. For example, “Village of Zermatt” (subject) is a photograph by (predicate) William England (object). Or “Village of Zermatt” (subject) was exhibited in (predicate) 1881 (object). Or “Village of Zermatt” (subject) was exhibited in (predicate) the Royal Photographic Society Annual Exhibitions (object). It is easy to see how such simple and unambiguous statements can be linked together to deduce for example that William England exhibited “Village of Zermatt” in the 1881 RPS exhibition. The goal of the W3C SWEO Linking Open Data community project is to accelerate the development of the Semantic Web by publishing various open data sets as RDF on the Web and by setting RDF links between data items from different data sources. This kind of approach is obviously relevant to GLAMs and recently there have been a number of efforts to publish metadata about the objects in GLAMs as linked Open Data (Batjargal et al., 2013; Kamura et al., 2011, Oommen and Baltussen, 2012). Notable examples include the Rijksmuseum (de Boer et al., 2012) and the Linked Open Data in Libraries Archives and Museums (LODLAM) initiative.

Promising as the API and Linked Data approaches are, there are large practical barriers to their widespread implementation. The GLAMs and Semantic Web communities have different terminology for similar metadata concepts, making implementation of semantic approaches difficult and while the Linked Data Web already comprises billions of RDF triples, the majority of Web documents are not yet in this format. Consequently, building a data pipeline to convert heterogeneous data from various source collections into RDF in order to create linked open will require considerable resources and technical expertise (Henry and Brown, 2012). While the onus is on GLAMs to standardize and convert their records to linked data the prospect of universally linked data remains remote because of the enormous cost of manually RDF-encoding billions of documents and the challenge of creating the appropriate external controlled vocabularies (Bizer et al, 2009). APIs remain equally challenging. Again there is a need for huge amounts of effort to manually convert existent GLAMs records into properly structured, syntax dependent XML.

Not only are the majority of records still encoded in non-machine-readable natural language that use a variety of different data structures and file types, but the data in the records themselves tend to be subject to a variety of quality issues:
• **Different metadata schemas** - different labels were used for essentially the same descriptor by different institutions, eg. ‘creator’, ‘artist/maker’, ‘photographer’, ‘exhibitor’, ‘auteurs’.

• **Incomplete records** - empty fields.

• **Junk data** - eg. mixed data fields such as person name and date of birth/death or date and place of birth in the same field, or date information expressed in text form: “circa pre Great War”. Again, while such information is intelligible to human beings, computers find it challenging when numerical and text data are combined or different types of text such as person and place names are bracketed.

• **Syntax independent formats** - eg. name order such as Henry Tomas Malby; H.T. Malby; Henry; T. Malby; Malby, Henry,Tomas, and so on. Made more difficult as formats are not used consistently within individual collections.

These issues currently prohibit the use of automated tools and processes to convert such records to machine-readable formats. Syntax independence is particularly challenging and until records are formatted consistently Linked Data will remain just another computing format rather than a gold standard. Yet the need to aggregate and compare records remains and is growing in importance as more and more collections are digitised. Without the ability to compare easily and effectively across different collections, much of the value of digitisation will remain unrealised. Therefore, until such time as a significant proportion of GLAMs records are improved, interim ways have to be found to deal with the messy reality of records as they currently stand. This is the ‘real world data’ challenge. The remainder of this paper describes an approach for helping researchers to find potential matches between such items in different collections that was devised to deal with the messiness of actual GLAM records.

2. The Exhibitions of the Royal Photographic Society

The starting point for this work was a previously published corpus of photographic catalogue records relating to the exhibitions of the Royal Photographic Society (RPS) between the years 1870 - when the exhibitions began - and 1915 - when international activities were constrained by the Great War. They contain over 45,000 records about people; exhibitions; lectures and exhibited photographs and photographic equipment from a
significant period in the history of photography; offering a unique insight into the evolution of photographic technologies; aesthetic trends and the activities of a flourishing group of companies responding to business opportunities; as well as the activities and fortunes of individuals who contributed to the technical, artistic and commercial development of photography. The Royal Photographic Society’s exhibitions attracted a wide range of photographers, from Britain, Europe and America and many individuals launched their photographic career through them. The catalogues provide details of exhibitors names, addresses, RPS membership status, exhibiting patterns, prize winning status, photographic exhibit titles, sale prices, photographic processes, exhibition categories and information about the exhibitions including membership of hanging and judging committees. Yet, as figure 1 shows, most exhibit records were text only. Out of 34,917 exhibits, only 1040 were illustrated and many of those illustrations were only artists’ sketches of the original photographs (see figure 2) because, at the time of the exhibitions, mechanical reproduction of photographic images was technically difficult and expensive and in many cases unnecessary, since the photographs themselves were on view in the exhibitions.

*Figure 1: Sample catalogue pages from the 1899 exhibition showing illustrations.*
*Source: http://erps.dmu.ac.uk*
Figure 2: Detail from the 1899 exhibition catalogue: Exhibit 76. A Scottish Loch.

RPS exhibition catalogues between 1870 and 1915 were digitised and published online in 2008\textsuperscript{11} as part of the broader corpus of photographic history resources developed and managed by De Montfort University\textsuperscript{12} and
have been well received by researchers and historians. However a recurring question from users has been “where are the pictures?” The “FuzzyPhoto” project described here aimed to develop computer based finding aids that could help researchers to answer questions such as: “How many of the pictures referred to in the exhibition catalogues have survived?”, “What do those pictures look like?”, “Where are they?”.

The approach entailed identifying collections in other institutions likely to contain examples of the missing images, ingesting the records from those institutions, cleaning the data and mapping it to a common metadata schema, mining the resulting data set for similarities between records and publishing information about any identified pairs back to the partner web sites for visitors to those sites to discover. Collections held by six GLAMs were identified as potentially useful and partnerships were established with these: Birmingham City Library; the British Library; the Metropolitan Museum of Art; Musée d’Orsay; the National Media Museum; St Andrews University. Additionally access to records from five other organisations was obtained during the project: the National Archives - National Museums Scotland; Brooklyn Museum; the US Library of Congress; Culture Grid.

3. Data cleaning

Having acquired the records, the intention was to import them into a common MySQL table formatted data warehouse ready for mining. However, it was immediately apparent that the quality of the records was too variable for this to be accomplished in a single step. Records ranged from highly structured and rigorously expressed syntax dependent XML through to all the data in a single column spreadsheet expressed in natural language and with no consistent formatting or syntax. Duplicates, empty fields and junk data were also common. An intermediary step was introduced to create a temporary MySQL database within which the records could be cleaned. CSV files were imported directly into MySQL, but for XML data, an XML data store (BaseX) and XQuery queries had to be used to convert data to MySQL tables, and an intermediate Microsoft SQL database was necessary to convert British Library data. After cleaning and removal of duplicates and corrupt records the total data set amounted to 1,406,666 records. Next, they were mapped to a common metadata schema and exported to the data warehouse ready for data mining. CIDOC-CRM was originally considered, but rejected in favour of LIDO because the partner metadata was so
heterogeneous, it would have been necessary to edit the all the records extensively to comply with CRM. The ICOM Lightweight Information Describing Objects schema (LIDO) was ideal for standardizing the metadata from each of the contributing organisations because it is an XML harvesting schema developed specifically for exposing, connecting and aggregating information about museum objects on the Web.

4. Finding matches

Four pieces of information were utilised from each record to identify matches: the title of the photograph, the name of the photographer, the photographic process used to create the photograph and the dates associated with it. The similarity metric used for processing the title fields was a customised short text similarity metric. Person names were analysed using a customised person name similarity algorithm derived from established edit distance techniques, combined with a heuristic best fit approach to match up individual name elements across fields, despite the absence of a standard name format. Photographic process information was analysed using a graph transversal algorithm to find the shortest path between the processes, based on a customized ontology of photographic processes. Dates were analysed using a custom developed date similarity algorithm that calculates the differences between different date spans. Finally, a fuzzy inferencing system was used to combine individual fields into an overall record similarity metric and to group the co-referent records together.

4.1. Title Field

Exhibit title records are unlike the kinds of text usually subjected to data mining in as much as titles do not follow normal grammatical rules for sentences and are generally brief. The average title length across all the exhibition records is just 8.1 words, of which only 5.4 are useful. This ruled out standard corpus analysis tools. While there are tools for analysing short texts, such as Latent Semantic Analysis (LSA) and Short Text Semantic Similarity (STSS), these cannot easily process large numbers of records (O'Shea et al, 2008) and the 1.4 million records we had to process required approximately $1 \times 10^{12}$ comparisons. It was necessary therefore to develop a customised semantic similarity measure capable of handling large numbers of very short records (Croft et al, 2013). This Lightweight Semantic
Similarity (LSS) tool is based on standard statistical cosine similarity metrics (Manning and Schütze, 2003) but additionally takes into account the semantic similarity between words, using WordNet, in combination with a minimum term similarity threshold.

4.2. Person Names

Name comparison is a common problem in many application areas and there are many well-established comparison algorithms that can handle typographical errors, alternative spellings etc. and even different syntax such as Firstname/Lastname, Lastname/Firstname, Lastname/Initial/Firstname, etc. We used an established edit distance technique (Jaro-Winkler) that measures similarity in terms of the number of changes (edits) that are required in order to convert one string into another (Winkler, 1990). However we had to combine this with a heuristic best fit approach because the person name syntax varied not only between collections but even within some of the collections, making the name order impossible to specify.

4.3. Photographic Process

Now, in an era of digital photography, it is difficult to imagine how many different photographic processes there were to choose from in the early years of photography.

Leaving aside complications such as creation of photographic negatives or photomechanical reproduction of images, there were around 40 different positive processes available, many of which were variations on each other and easily confused. Misattribution or uncertain attribution of process are common issues in historical photographic records, with obvious implications for this study: Two otherwise similar photographs mistakenly attributed to different photographic processes may in fact be identical. To accommodate this, the various photographic processes listed in the records were organised into an ontology in which processes sharing specific traits appeared close together. Once the field was matched to a specific locus it was compared to other fields using a graph transversal algorithm to find the shortest path between the processes. The shorter the distance between the approaches, the more similar they are considered to be.

4.4. Dates
Date information in the records was used to assess similarity, based on the logic that records with dates close to each other are more similar than records with dates further apart. Unfortunately the syntax of date information tends to be highly variable. For example, some dates are ‘day/month/year’, others are ‘year/month/day’ (one institution managed to find 20 different ways of recording date information). Many of the dates are also imprecise, such as ‘the 1890s’, ‘the 19th century’, ‘circa 1870’. Extraction of the date information from the various formats used was achieved with a combination of the Python dateutil library, regexes and a rule based system whereby greater differences between time spans indicate less similarity between fields. For example ‘19th century’ and ‘1900’ are less similar than ‘1900s’ and ‘1900’.

4.5. Combined Similarity Score

To arrive at an overall similarity score for each pair of records it was necessary to find a way of combining the individual field metrics. Probabilistic Record Linkage (PRL) was considered, using manually tuned probability weightings but early experiments showed that as a naive-bayes classifier, PRL was unsuitable. Artificial Neural Networks (ANNs) were similarly considered and discarded, because the training data set requirements would have been impossible to meet.\textsuperscript{15} To deal with the imprecise and uncertain information inherent in the records and hence the individual field metrics, a fuzzy inferencing system was deployed, based on fuzzy logic. Fuzzy logic has been applied previously to resource discovery challenges (Feng, 2012; Lai et al., 2011; Li et al., 2009). However, as these approaches were based on analysis of large volumes of text it was necessary for us to develop our own rules arrived at through a combination of trial and error and a survey of members of the GLAM community regarding the relative importance of the different variables:

- If bad_title AND bad_person THEN terrible_match.
- If bad_title OR bad_person THEN bad_match.
- If good_title OR good_person THEN good_match.
- If good_process AND good_date THEN good_match.
- If good_title AND good_person THEN excellent_match.

Finally, each resulting fuzzy set was converted into a crisp value through a
process called defuzzification, described fully in Coupland et al. (2014). These values were used to populate a series of dendrograms, starting with a seed record at the root node. FuzzyPhoto identified the records with the highest similarity to the seed record and exceeding a pre-set threshold. Additional levels of child nodes were added until the dendrogram reached a maximum size or the similarity values dropped below the threshold. The end result is that those records with the greatest similarity to the seed record appear in the highest layers of the tree and connections can be inferred via intermediary records even where there is no apparent direct relationship.

5. Results

The project set out to find the pictures ‘missing’ from the ERPS catalogues. Figure 3 presents a snapshot (based on a sample of 50 records) of the distribution of matches discovered between ERPS records and those in partner collections, matched by person name only, by title field only and by combined person, title, date and process, labelled as ‘balanced’. Not surprisingly, most of the person name matches found are excellent, since names are usually unambiguous and have a limited number of possible representations. Matches based solely on title contain more uncertainty so the proportion of excellent matches is lower. Inevitably there are fewer ‘balanced’ or overall matches based on the combination of all four fields simply because increasing the number of fields included in the comparison increases the opportunity for records to differ from each other.

Figure 3: Number of records in the test sample demonstrating at least one balanced/title/person match that is excellent/good/possible. \(^{16}\)
Figure 4: Proportion of balanced/title/person matches in the test sample rated excellent/good/possible.\textsuperscript{17}

Figure 4 presents the same results, this time as proportions of the sample rather than as absolute numbers, providing an indication of the possible distribution of matches across the total data set. From figure 4 it can be seen that around 10 per cent of the sample appear to be excellent matches.

In this context excellent does not mean identical. The threshold for excellence has been set low enough to ensure that potentially good matches are not excluded accidentally. However, in a number of instances the match is so close we can be reasonably confident that the missing image has been rediscovered. Figures 5 and 6 show two examples of excellent matches between records in the RPS exhibition catalogues and records discovered in partner collections. Since the RPS records are not illustrated, we cannot be absolutely certain that these records co-reference the same photographs, nevertheless the high degree of correspondence between them suggests that the photographs illustrated here are indeed the same as those exhibited at the RPS, as listed in the catalogues.

Figure 5: Example of a likely exact match with an ERPS record.
Figure 6: Another example of a likely exact match with an ERPS record.
6. Performance

In order to assess how the co-reference identification speed and accuracy of FuzzyPhoto compares with human experts, we tested a sample of the outcomes on a panel of subject experts. The results indicate that FuzzyPhoto finds more matches than are found manually and the matches suggested by FuzzyPhoto are at least as good as those discovered by experts. However an issue that arose during these trials concerns the way that some researchers interpret the notion of a match. Some defined a match as some other picture by the same photographer, while others understood it to mean something with a similar title or subject matter, or made by a similar photographic process, even if the photographer name was different. Accordingly, the Web interface was adapted as shown in figure 7 to provide users with a choice between searching for similarity by person name, title or by overall similarity based on the combined metric.

Figure 7: 7 Screen shot of the final version of the FuzzyPhoto widget embedded in the ERPs Web site, opened to show links categorised by person name, object title, or overall similarity (all fields).
7. Discussion

As the number and volume of online museum collections grows there is an increasing imperative to improve their discoverability by finding ways of linking records that go beyond simple keyword searching. Keyword searches are inefficient because they are prone to errors of both omission and commission. A range of more sophisticated approaches have been developed for cross collection searching including metadata harvesting, data mining, Linked Data and Application Programming Interfaces (APIs) but these variously rely on availability of well structured, consistent and standardised data and a large corpus of text. While some GLAMs records meet these standards, many do not since they employ different data schemas, applied inconsistently, are often fragmentary and imprecise and non-machine readable. Thus, while some heritage institutions have successfully
implemented one or more of these computational approaches, pioneering the way for others to follow, the majority of online collection records are not amenable to such treatments. A further challenge at the core of this study is that there are many millions of individual object records, each containing many different fields, the entries in each separate field tend to be quite small, e.g. dates, person names, titles, making it difficult to apply corpus based approaches to data analysis. The cost of converting fragmentary, messy museum records into well-structured data is too great for many institutions faced with the challenge of keeping up with rapidly growing collections, let alone reworking their back catalogue, even if the necessary skills were in place. So while linked data and APIs are useful in some cases, attempts to link records across the majority of institutions will have to deal with the messy reality of GLAMs data for the foreseeable future.

In this paper we have described a fuzzy inferencing system combined with a constrained hierarchical clustering approach that can accommodate messy real world data using multiple metrics tuned to the specific challenges of the information held in specific fields, combined into an overall record similarity metric with the aid of fuzzy inferencing. Using this approach we have successfully identified matches for around half the records in the catalogues of the Royal Photographic Society annual exhibitions, allowing images of some of the exhibits to be seen for the first time in over 120 years. In addition FuzzyPhoto has identified matches between partner records, even when there are no matches with exhibits in the RPS exhibitions. So, for example, photographs in the Library of Congress have been matched with similar items in the British Library, and items from the National Media Museum have been linked to photographs in the Musée d’Orsay. Trial results indicate that this approach is at least as effective as expert researchers in identifying potential matches between records and considerably faster. This suggest that FuzzyPhoto is useful not only for rediscovering the lost images from the RPS exhibitions but can be applied more generally to the task of identifying potential connections between collections across a range of GLAMs institutions, enriching our understanding through cross-referral and adding value to the records already created.

8. Acknowledgements

This research was supported by the UK Arts and Humanities Research Council [Research Grant AH/J004367/1]. Thanks are also due to colleagues
from Birmingham City Library; the British Library; Musée du Louvre; the Metropolitan Museum of Art; Musée d’Orsay; the National Archives; the National Media Museum; National Museums Scotland; St Andrews University; and the V&A for their generous support.

References


Brown, Stephen, David Croft and Simon Coupland. 'Methods for Mining Messy Real World Data: Co-reference Identification Using Fuzzy Logic'. Source: https://www.hrionline.ac.uk/openbook/chapter/dhc2014-brown


January 2015).


Footnotes


10. 1914-1918.


15. The required training data set would have had to contain matched pairs of records, which is precisely what this study was attempting to produce.

16. Based on a sample of 50 records.

17. Based on a sample of 50 records.

18. Based on analysis of a random sample of the full data set.