FUZZY PHOTO PROJECT

Sarah Greenfield

28th September 2010

Contents

Chapter 1

The DMU Online Photographic Databases

1.1 The Problem

The De Montfort University (DMU) photographic research databases of historical photographic exhibitions are popular with researchers but, researchers frequently ask, “Where are the pictures to illustrate the photographs referred to in the exhibition catalogues?” Increasingly digital pictures of historical photographs are available online via galleries and archives, but now the question is, “Which ones match the catalogue records in the DMU photographic research databases?”

The DMU exhibition catalogue records information about features such as exhibitor, exhibit title, date (of exhibition), and (photographic) process. Online archives of pictures usually include similar kinds of information such as names and dates — similar but not necessarily the same. For example, names of people in photographic picture archives are usually of the photographers rather than exhibitors, so a direct match may not be possible because while in many cases the person exhibiting a photograph was the person who took the photograph, this was not always the case. Moreover, people’s names and titles can change, for example, on marriage. Picture titles are similarly unreliable. The same picture may be known by a variety of different titles. So the process of matching available pictures to the records of pictures in exhibition catalogues is not entirely straightforward.

An approach is needed that individually compares each picture and its associated record with the historical catalogue data and reaches a decision as to the likely degree of ‘fit’. While this could be done manually by experts, such an approach would be very time-consuming and expensive. An alternative approach would be to develop some form of computational algorithm for making such comparisons and suggesting possible matches. The matching algorithm is intended to be be pre-run against the available online collection databases, and the results stored.

1.1.1 The Research Question

This is the research question addressed by the Fuzzy Photo Project:

What is the best computational method for identifying such matches reliably and efficiently?

Four subsidiary questions arise out of the research question:

1. **Is fuzzy logic a suitable technique for finding matches? Or is string matching sufficient on its own?**

2. **What is the ideal metadata schema to support matching records?**

3. **What type of database system is most appropriate for matching — a conventional relational database system, or an XML/RDF database system?**

4. **Which type of database system is most suitable for storing the matches — a relational database system, or an XML/RDF database system?**

STEPHEN BROWN: So we're not trying to create an artificial intelligence system here like say a medical diagnosis system. It's just . . . a filtering system which gives people a short list from which to work, and so they can now say well, 'It might actually be worth going to Washington to look at that collection now, or trying to get access to the Royal Collection at Windsor or go to Kew and look at all the photographs there. I know now where it's worth starting to look, as opposed to start with *A* and work through to *Z* until I find something.'

1.1.2 Conventions Adopted Within the Report

1. *Fuzzy Photo Matcher* (FPM) is the working name for the planned software implementation of the Fuzzy Photo Project recommendations.
2. Where quotations contain citations, the citation number has been adjusted to correspond to the report references.

1.1.3 Assumptions

1. The FPM only applies to databases implemented in English. (The databases may contain foreign words as data, but the field names and metadata must be English.)
2. All data is recorded using the English alphabet.

1.2 Metaphysical Issues Associated with Photography

Let us imagine a person called John Smith who writes crime thrillers under the pen name Morris Mitchell. Both names refer to the same person, though this is not commonly known. Those in the know would describe John Smith and Morris Mitchell as being *the same* person, reflecting the fact that there really is only one person, but he has two names. Now let us imagine that John Smith has an identical twin brother called James Smith. They are so alike that they are virtually impossible to tell apart, and are frequently mistaken for one other. Those who know them both say that they are *the same*. By this they mean that they are indistinguishable — not that there is really only one person. When mistaken for each other, John and James are able to correct the error simply by stating their real identity. What this example show is that the word 'same' can can be used in the sense of identity (where there is one), but also in the sense of indistinguishability (where there is more than one).

Two prints from the same negative will be indistinguishable, provided they have been through exactly the same development process[1]. Suppose the FPM establishes a match between a catalogue entry in one of the DMU databases, and a photograph in an online collection. Is it the actual photograph that was exhibited, or is it another photograph, identical to the exhibited one? It could be that a number has been written on the back of the photograph which corresponds to the catalogue number of the exhibition. That would indicate that it is the actual photograph exhibited. But if there is no corroborating evidence, we shall never know for sure; unlike John and James Smith, the photographs cannot speak for themselves[2].

There are further complications. A negative may be used to create prints which look slightly different by altering the parameters of the development process, e.g. increasing the time the print is immersed in developer fluid. The two resultant prints would be described as '*almost the same*', though some might say that the *image* was *the same* (subsection 4.3.1). This would be analogous to John and James Smith deliberately making themselves distinguishable by colouring their hair differently.

[1] Indeed two prints from different negatives taken under identical circumstances will be indistinguishable.

[2] To a researcher of photographic history, it would still be a useful match.

KELLEY WILDER: . . . it might be that that artist made 20 prints of it but only one of them is the one that was actually shown.

SARAH GREENFIELD: And they can be slightly different, can't they?

KELLEY WILDER: Well each print was a piece of work in and of itself — they're very handcrafted items. . . . They're not wholly different; they're all made from the same negative so they probably look fairly similar, but they're very often not *the* one that might have been shown. Even if you found the image, you're still not finding *the one* that might have been shown — you're finding the right image that corresponds to one that was shown, but maybe not the actual artifact, the actual object that hung on the wall.

SARAH GREENFIELD: That would still be relevant though, wouldn't it — it would . . . be useful?

KELLEY WILDER: . . . it would be very useful, but I think it's important to keep the wording careful. 'Cos people do tend, in photography, to mix up, or to forget, that . . . each material object has its history, the image itself has a history in image culture of being shown in various places, but you can actually have the same image hanging in two places, but not the same object hanging in two places at the same time.

SARAH GREENFIELD: Presumably you could actually not know whether it's the actual one.

KELLEY WILDER: You probably would never know. . . . It's something you just have to live with.

When we start attaching names to prints, it becomes even more confusing. A photographer may take several shots of the same scene with minute variations of the camera angle or of the scene itself. The resulting prints would be hard to tell apart, and again, would be regarded as *almost the same*. However these photographs may have all been given the same name in the original exhibition (or in an online collection). For example, *Mechanic*, in the George Eastman House Museum [16], is actually a group of five slightly different photographs, all sharing the same title.

The DMU databases have instances whereby several photographs are recorded as one exhibit. For example, exhibit 15, *Sea Views*, documented on page 2 of the 1883 ERPS catalogue, is actually a group of nine pictures. The nine pictures have individual titles, some of which are shared, e.g. pictures 7, 8 and 9 are all called 'Sea Larks'.

There are instances where it is impossible to know which one image is signified by the catalogue entry, as a photographer may have given the same name to several similar images (which have not been grouped together as in the *Mechanic* or *Sea Views* examples):

KELLEY WILDER: . . . and one artist may have taken 10 views of Tintern Abbey and called them all 'View of Tintern Abbey', and you may not know which one specifically that relates to. Very few people titled their images so carefully as to know which one was which, and they didn't tend to number them . . . , 'Abstract 1', 'Abstract 2', as they do now!

A print may have been exhibited in an exhibition under one title (name), later appearing in an online collection under a different title, with different attributions of key people such as owner, exhibitor etc.. Most of this variability is due to the fact that photographs have changed hands and been renamed. However mistakes may have been made in the documentation and there is even the possibility of fraud.

Chapter 2

The DMU Photographic Research Database Data

2.1 The Computer Scientists' View of the Data

2.1.1 Data Types

Familiarisation with the DMU photographic research databases revealed that from a computing point of view there are four data types to be found, namely *string*, *numerical*, *categorical* and *Boolean*.

String A string is simply a list of characters such as a word, phrase, or sentence. Examples of string data within the DMU databases are *exhibit title*, *exhibitor name*, *exhibitor address*, and *'photograph by'*.

Numerical Exhibition *dates* are numerical. More specifically, since the dates are the catalogue years, they are *integers*[1].

Categorical Data that falls into categories is termed *categorical*. Examples of categorical data are *exhibition society*, *exhibitor title*, *exhibitor qualifications and affiliations*, *RPS membership status*, and *process*.

Boolean Boolean data is a special case of categorical data, where 1. there are only two categories, and 2. each datum is obliged to belong to one category or the other. Examples of Boolean data are *exhibit medal status* and *exhibit picture in catalogue status*.

2.1.2 Data Fields

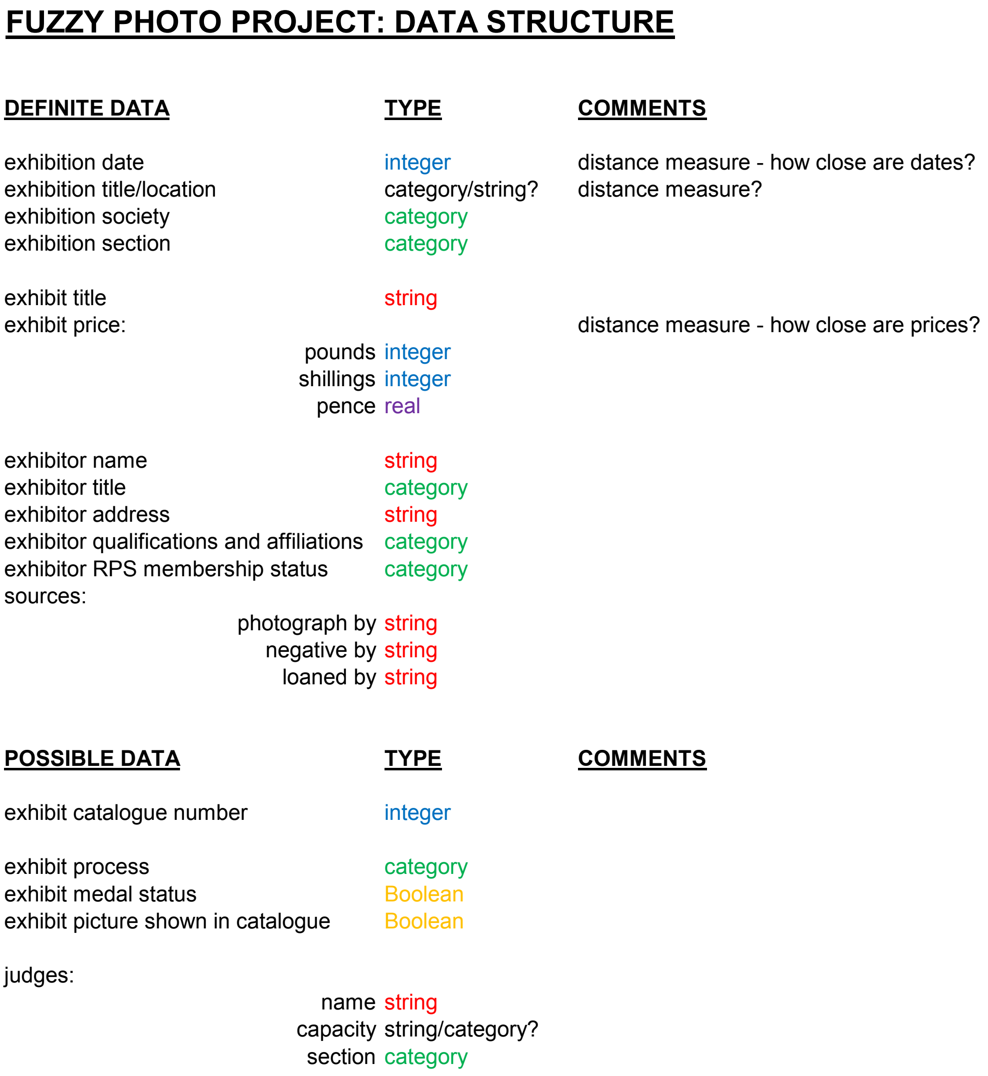
Figure 2.1 is a spreadsheet showing all the data fields in the DMU research databases. The fields were split into those thought useful (DEFINITE DATA) for photograph matching, and those not (POSSIBLE DATA).

2.2 The Photographic Historians' View of the Data

2.2.1 Knowledge Elicitation Exercise

The Fuzzy Photo Project was envisaged as being realised by a Fuzzy Inferencing System, hence its title. However we had to consider the possibility that photograph matching might be better achieved through string matching alone (subsection 1.1.1). To decide between a fuzzy and non-fuzzy system we needed to learn the inferencing rules (section A.4) of the matching process. As we did not have the historical data to underpin a data driven approach, a *knowledge elicitation exercise* was undertaken.

[1] An integer is a whole number.



FUZZY PHOTO PROJECT: DATA STRUCTURE

DEFINITE DATA	TYPE	COMMENTS
exhibition date	integer	distance measure - how close are dates?
exhibition title/location	category/string?	distance measure?
exhibition society	category	
exhibition section	category	
exhibit title	string	
exhibit price:		distance measure - how close are prices?
pounds	integer	
shillings	integer	
pence	real	
exhibitor name	string	
exhibitor title	category	
exhibitor address	string	
exhibitor qualifications and affiliations	category	
exhibitor RPS membership status	category	
sources:		
photograph by	string	
negative by	string	
loaned by	string	

POSSIBLE DATA	TYPE	COMMENTS
exhibit catalogue number	integer	
exhibit process	category	
exhibit medal status	Boolean	
exhibit picture shown in catalogue	Boolean	
judges:		
name	string	
capacity	string/category?	
section	category	

Figure 2.1: Initial data classification.

Knowledge elicitation took the form of expert interviews. The first interview was of Stephen Brown, regarded as a 'semi-expert' in the field of photograph matching. Following that, three other experts were interviewed:

Dr. Kelley Wilder Senior Research Fellow at De Montfort University. Programme Leader of the MA Photographic History and Practice at DMU, and a member of the Photographic History Research Centre.

Dr. Jane Fletcher Senior Research Fellow at De Montfort University, module leader on the MA Photographic History and Practice at DMU, and a member of the Photographic History Research Centre. Former RPS Collection curator at the National Media Museum in Bradford.

Mr. Michael Pritchard Research Fellow at De Montfort University, module leader on the MA Photographic History and Practice at DMU, and a member of the Photographic History Research Centre.

As well as shedding light on the experts' matching strategies, the knowledge elicitation exercise (from which the quotations are taken) was most enlightening about the *nature* of the data. The source data of both the DMU catalogues and the online collections is subject to inaccuracies and incompleteness. On the DMU side, the inaccuracies have largely been addressed.

> JANE FLETCHER: The problem is ... that the source material isn't necessarily correct. So if material's been gleaned from the journals, names get spelt wrong, ... things are omitted. ... there's the possibility for data to have been inputted correctly but the source is wrong initially. But I imagine that has been addressed in the database ... a lot of cross-referencing went on to ensure that spelling of names was consistent. So I think that mostly has been addressed.

The DMU data has been corrected as far as possible. However we have to assume the online collection data to be unreliable.

2.3 Inconsistencies in the String Data

From the outset it was clear that string matching would constitute part, if not all, of the inferencing process. The string data is rife with inconsistencies, which fall into two groups: 1. inconsistencies of wording, and 2. inconsistencies of spelling.

2.3.1 Inconsistencies of Wording

Names may be inconsistent for three reasons:

Promotion People's names alter through promotion. However, even in the extreme example of Sir Charles Abney, 'Abney' was always *part* of his name.

Marriage Marriage changes a woman's surname, *but her first names remain the same*.

Renaming For historical and political reasons places may have more than one name.

When addresses are inconsistent it is usually because people have moved house.

> STEPHEN BROWN: When we get into things like exhibitor title and exhibitor address, or indeed exhibitor qualifications and affiliations, these are things that just help to qualify that exhibitor name, because names do change — the classic being 'Abney' [who] appears as 'C. E. Abney', 'Lieutenant Abney', 'Captain Abney', 'Sir Charles Abney'. He did well! And then of course you have women who've married and then their name changes completely.

> KELLEY WILDER: In some cases some exhibitors have two or three names that all mean the same exhibitor.

Exhibitors are mostly people but can be organisations such as businesses. The formatting of names is variable; initials and other abbreviations are used inconsistently, e.g. 'Berlin Photographic Company' and 'Berlin Photographic Co.' both appear as exhibitors in ERPS. A couple can be recorded as one exhibitor, e.g. 'Beck, R. & J.' in ERPS. Some cryptic names from the PEIB database are: 'A Boy 12 Years of Age', 'A Lady', and 'Amateur'.

SARAH GREENFIELD: Is there anything else that you just wouldn't trust very much?
KELLEY WILDER: Well, addresses, because they change so often …

Addresses are often incomplete.

MICHAEL PRITCHARD: exhibitor addresses, …they can be very sketchy, so it could just say, 'London', which isn't overly helpful — it's not going to be, '22, Acacia Avenue'. …On some of the exhibition catalogues they're very sketchy, …as you progress through the century, then sometimes they get better.

The RPS membership status is also liable to change.

JANE FLETCHER: Obviously the RPS membership status will change as well. …People go from associate to honorary …There's various kind of status that change.
SARAH GREENFIELD: Have you got to be a member to exhibit?
JANE FLETCHER: Yes, I believe so. …Prior to 1974 or something there's the associate and the fellowship of the RPS, and that follows the name as ARPS or FRPS, so that obviously changes along with addresses.

Titles are very varied indeed. A verse of poetry might be used as a title. For example

> "Soon as the silent shades of night withdrew,
> The ruddy morn disclosed at once to view
> The face of Nature in a rich disguise'
> And brightened every object to my eyes."

is the title of exhibit 54 in the 1883 ERPS catalogue. Titles can also be extremely unspecific. There are numerous instances of 'Study', 'Portrait' and 'Landscape'. Titles might be fairly specific, yet in a foreign language, e.g. exhibit 140 in the ERPS 1908 catalogue has a German title, 'Schlafzimmer im Schloss', which translates to 'Bedroom in the Castle'.[2]

MICHAEL PRITCHARD: Exhibit title …they're not always used consistently so when you see what's likely to be the same image in different catalogues, but you're never quite sure because it's a group of trees, or it's given a more poetic title but it's almost certainly the same image.

Where titles refer to place names, there is no guarantee that it can be found on a modern map, or even a map of the relevant period!

KELLEY WILDER: There are a lot of places on the Indian subcontinent that either no longer exist, or were under a different name. Or for that matter, there are a lot of places in Europe that had different names, and that is fairly common — you find that quite a lot. And of course the *Constantinople–Istanbul* problem. That's another one that has to be watched with care 'cos it changes a lot [in that period of history]. And across the two databases, of course, because you're talking a half century, really, so it is quite a long time, really, historically speaking, and place names tend to just shift, even in Britain where they were known as two different things and then one becomes standardised as they begin to standardise maps and do Ordnance Survey and that kind of thing …

[2] This is interesting, but not a problem for the Fuzzy Photo Matcher, which will match strings in any language provided the English alphabet is used.

2.3.2 Inconsistencies of Spelling

With every written word there is potential for a variant spelling. Spelling inconsistency may arise from three sources:

Alternative Spellings Some words have alternative spellings owing to spellings not having been standardised. This is especially so with names.

Misspellings Words are often misspelt.

Typographical Errors "A **typographical error** (often shortened to **typo**) is a mistake made in, originally, the manual type-setting (typography) of printed material, or more recently, the typing process."[3] [56]

MICHAEL PRITCHARD: Exhibit title — again there's misspellings ...

KELLEY WILDER: In the metadata, of course, under exhibitor name and title, there are obviously several entries for some of them because of spelling discrepancies. ... Some places have different spellings, like popular places in Wales often were spelt in different ways, copies of Michelangelo paintings — 'Michelangelo' was spelt in very odd ways. ... 'Herschel' with one *l* and 'Herschell' with two *l*s. There were ... several place names in Wales that are in titles that we had to account for. 'Michelangelo' was a big problem, of being in titles. I have a feeling that there were two or three other names that we had to deal with, but I can't recall them off the top of my head. ... It's common because it's nineteenth century so spellings were not as standardised ... — a lot of place names spellings especially, it seems. ... Many things too, which were hyphenated, you'll find by the end of 20 or 30 years, they'll have become one word instead of being hyphenated, and that was the case with Michelangelo ... I think what we found was there was 'Michel–Angelo', 'Michel Angelo' and 'Michelangelo' — all three. ...

SARAH GREENFIELD: So did you arrange it so that if you searched on one you'd get the others?

KELLEY WILDER: Yes.

SARAH GREENFIELD: We can't do that because we're dealing with other people's data.

KELLEY WILDER: ... it strikes me is on the other end you'll be having the same problems.

SARAH GREENFIELD: Absolutely.

MICHAEL PRITCHARD: ... even names, someone like Delamott, for example, it's 'Delamott' as a single word, it's 'De la Mott', it's 'De la', it's 'De le', so there's certainly a lack of consistency there.[4]

2.3.3 Processes Data

The drop-down search menu offers the researcher a choice of six processes:

- Colour,
- Negative,
- Photomechanical,
- Positive,
- Print, and
- Printing.

These six categories are umbrella terms for processes such as *Gelatine plates, own make* (15), *Woodbury enlargement* (17), and *Platinotype* (35).[5] Processes are strongly associated with photographers.

[3] Both the DMU research databases and the online collections are susceptible to typographical errors in data input.

[4] There are many examples of photographs by Philip Henry Delamotte (1821-1889) in the PEIB database.

[5] ERPS webpage "Catalogue page 2 of 28 from the Exhibition Catalogue 1883 [Twenty-eighth] Photographic Society of Great Britain Exhibition".

STEPHEN BROWN: Certain people are associated with certain processes.

The drop-down menu presentation gives the initial impression of crisp, clear category data, but the expert interviews reveal that is far from being the case. The process data, though important, is highly unreliable.

KELLEY WILDER: It's unreliable in that not everyone described process the same way. Some people described the negative process and some people described the printing process as *the process*. ... Process description turned out to be a very slippery category. ... Many processes had many different names. ... Some processes are now known as other processes, some of them are just mild variations on one another, some of them are actually the same but used in the States or in Britain or in France. Sometimes we don't really know what all those mean.

MICHAEL PRITCHARD: ... processes is probably the area, particularly in the 1850s and 1860s, ... [where] things do get misidentified even back at the time.

SARAH GREENFIELD: Leaving aside spellings and things like that, which of the actual data would you regard as least trustworthy — that's to say people wrote things down and they didn't really always know what they were talking about?

JANE FLETCHER: Well process is quite a problem. ... process is the main issue.

To complicate matters further, the same image may be produced using different processes.

JANE FLETCHER: Going back to the process, the only thing that can also become problematic is if a photographer has made the same image using different processes, which can happen. So someone like Julia Margaret Cameron might make the same image as an albumen print and also a carbon print. So there is slippage there. ... also you get examples where an original ... is a salted paper print, then years later, using the right negative a different print is made by a different photographer of the same image, who's effectively found the negatives and reprinted them as his own. So there are these slippages around. And the way we used to get round it at the museum is we used to talk about whether it was a vintage print, whether it was an original print, whether it was a copy from the right negative. But there are these curious mutations that can happen because somebody, a different photographer, uses original negatives to make prints in a different process.

The untrustworthiness of process data varies according to category; some categories are more trustworthy than others [Kelley Wilder].

Fortunately, fuzzy logic is well suited to dealing with unreliable data.

SARAH GREENFIELD: Well, fuzzy logic, especially type-2 fuzzy logic, has the answer to this. ... you can have an important piece of information with great uncertainty attached to it.

KELLEY WILDER: Ooh yeay, that's what that is, that's a perfect description of the process.

2.3.4 Dates

Dates in the DMU databases are given as years, and are reliable, as they are simply the year of the catalogue. Dates in the online collections are not necessarily the dates of exhibition (and anyway, the same print might be exhibited in more than one year at different PEIB exhibitions.)

A date can provide useful back-up information for matching or ruling out a match:

STEPHEN BROWN: Having a piece of date information might enable you to distinguish between **that** *Still Life with Fruit and Flowers* and **this** *Still Life with Fruit and Flowers*.

They may refer to the date of creation, and might well be estimates.

MICHAEL PRITCHARD: ... a lot of collections record with just a circa date ... that then becomes much more difficult to match or have some degree of certainty because your assessment of that circa date, 'circa 1870', — I might say '1865', and it could be 1860 in any case ... The online collections, I think that's where there's some real issues over the quality of the data that those institutions are putting on board and I think the date is one area where there's a lot of guestimates being put on by people who perhaps don't have the experience or knowledge to come up with something more definite. In some cases, if there's a stamp on the back of the print, or an exhibition label then it's going to be very clear-cut, but there's an awful lot of material that doesn't have that.

Dates have to be quite close to corroborate a match:

MICHAEL PRITCHARD: ... over a 3 year period, perhaps — that's the level of confidence I would have that it might be the same image being exhibited across multiple exhibitions. I think that by the time you get to ten years it's probably a different photograph. In my experience I don't think photographs got exhibited over such long periods until more recently when they'd come back as retrospectives and those sort of things. I think at the time, for a photograph to go across even two years, that's about the limit. ... to me, 2 years, perhaps 3 at a push, is the limit of the confidence I would have over that sort of dating.

SARAH GREENFIELD: How close a match would there have to be for it [the date] to make any difference in your view?

JANE FLETCHER: I would say 5 years is reasonable — if you were putting circa 1870 or something, I think you could go 5 years either way.

Dates do not necessarily appear in the online archives; when they do, they are within the description field.

2.4 The Data in the Online Catalogues

2.4.1 The Description Field

The description field in the online databases contains an assortment of information about the photograph. In this field will be found dates, catalogue numbers, print sizes, processes etc..

SARAH GREENFIELD: In what you said you imply that you expect the online catalogues to have records of the exhibitions. Is that generally your experience?

MICHAEL PRITCHARD: I wouldn't say generally; they sometimes can do, and again this is particularly later in the century — the early ones much less so, but by the time you get to the 1890s, 1900s then you start to get stickers and stamps being put on the back of prints, so depending on where those prints have come from into that collection, then they may well carry those sorts of records.

SARAH GREENFIELD: And that would be in a description, would it?

MICHAEL PRITCHARD: I would hope so. ... In fairness to the museums and galleries that have that sort of material, generally if they've got a reasonable cataloguing program will put that supporting information into their descriptions. So hopefully it should be online, but I wouldn't like to say there's consistency over that and a lot of descriptions are very sketchy. They just don't have the time or resources to do what I would call a proper cataloguing job on a particular image.

Exhibition stamps (on the backs of photographs) became more prevalent over time:

KELLEY WILDER: I know that 20th century items really do have exhibition stamps on the back, but I don't know when they started doing that.

The description field is very flexible and subjective in what it may contain.

MICHAEL PRITCHARD: The . . . exhibit description is probably the key piece there, because I think that's actually very subjective, from a person cataloguing the item in the 1850s or someone later. I think the remainder are much more definite types of data; pieces of data that should be correct if someone's being diligent in terms of the work they're doing with the items either back in the 1850s when they're putting them into a catalogue, or in a collection currently . . . they're cataloguing it for making it available online on a database.

JANE FLETCHER: . . . then we would have the exhibit title, the exhibit date, the exhibit process and the exhibit dimensions which would go under description, I presume — if they were available. Again it might be that the actual dimensions aren't available.

As with other data, descriptions are not entirely to be relied upon.

SARAH GREENFIELD: Leaving aside spellings and things like that, which of the actual data would you regard as least trustworthy — that's to say people wrote things down and they didn't really always know what they were talking about?

JANE FLETCHER: Exhibit description . . . I mean descriptions can get kind of simplified and changed through reporting and things.

As well as the description field, there may be the museum *notes* field available for searching:

KELLEY WILDER: . . . in the *notes* section, which is of course searchable, of museum entries, catalogue entries, that 'so-and-so won a medal for something', and they've actually done the leg work for you. . . . I would trust them more than processes to tell you the truth, if I had to rank them, but that doesn't mean that they can't be wrong.

2.5 Data Embedded in Titles

In both the DMU and online databases, a location or a (specifically named) person may be part of an exhibit title. The location may be a building, landmark, or population centre. Some examples of locations as subjects are:

- In 1888 the Autotype Company exhibited a photograph called, 'The Library, Beau Manoir, Leicestershire' (exhibit number 2).
- In 1888 G. C. Butler exhibited a photograph called, 'Brassworker's Shop, Muttra, India' (exhibit number 18).
- In 1888 G. C. Butler exhibited a photograph called, 'Shahpier, India' (exhibit number 19).
- In the 1888 catalogue, W. Hayes is recorded as exhibiting a photograph entitled, 'Kaynance Cove' (exhibit number 11).

Some examples of people as subjects are:

- In 1874 Spencer, Sawyer, Bird, & Co. exhibited a portrait of H. R. H. The Duke of Edinburgh (exhibit number 70).
- In 1913 Madame D'Ora exhibited a portrait of Comtesse Finette Wydenbruck (exhibit number 7).
- In 1913 Madame D'Ora exhibited a portrait of Professor F. Schimetrer (exhibit number 8).

Chapter 3

Matching Records to Photographs

3.1 Knowledge Elicitation Exercise

Prior to the knowledge elicitation interviews, the data fields were given a preliminary classification as either relevant or irrelevant to the matching process (figure 2.1). The experts were asked to

1. check that all the data fields were represented,
2. reclassify any data fields they believed to be wrongly classified (relevant/irrelevant), and
3. assess the importance of each piece of relevant data to the matching exercise.

The order of the interviews was 1. Stephen Brown, 2. Kelley Wilder, 3. Jane Fletcher, and 4. Michael Pritchard. Following the interview with Stephen Brown, the relevant/irrrelevant classification was amended slightly to give the final format which was kept constant for the the other three interviews (figure 3.1)[1]. There was an overwhelming consensus among the experts as to which data was relevant, and which irrelevant; any disagreements were relatively minor and unresolved.

- Kelley Wilder and Jane Fletcher saw the exhibition catalogue number as relevant.

- Michael Pritchard felt that the *exhibition medal status* was irrelevant, and questioned the relevance of *exhibitor qualifications and affiliations*, and *exhibitor RPS membership status*. He also saw *exhibition society* as relevant and *exhibition title/location* as important within the relevant data.

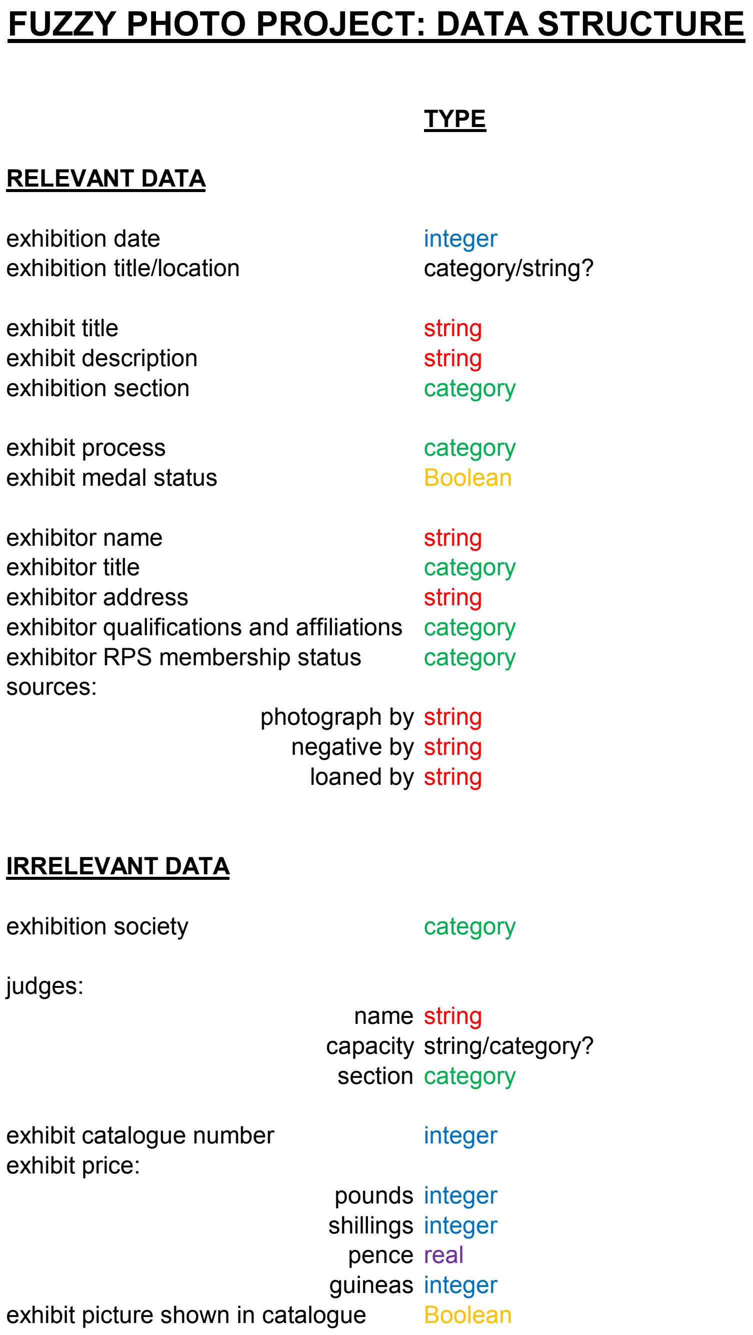
- Jane Fletcher saw the *exhibit picture shown in catalogue* status as relevant, if it were available in the online catalogue. Kelley Wilder saw this piece of information as absolutely crucial to the human matcher; if the photograph were known to be in the catalogue, it would obviate the need for the Fuzzy Photo Matcher. Whether desirable or not, it is difficult to see how this piece of information might be incorporated into the FPM.

The relevant data that they regarded as important fell almost entirely into four groups:

1. exhibit data,
2. historical stakeholder,
3. exhibit process, and
4. exhibition date.

Generally exhibit data and historical stakeholder data were seen as pre-eminent, with processes and date data corroborative. However, Michael Pritchard's view differed somewhat (figure 3.2). Relevant data of lesser importance was: 1. exhibition title/location, and 2. exhibition medal status.

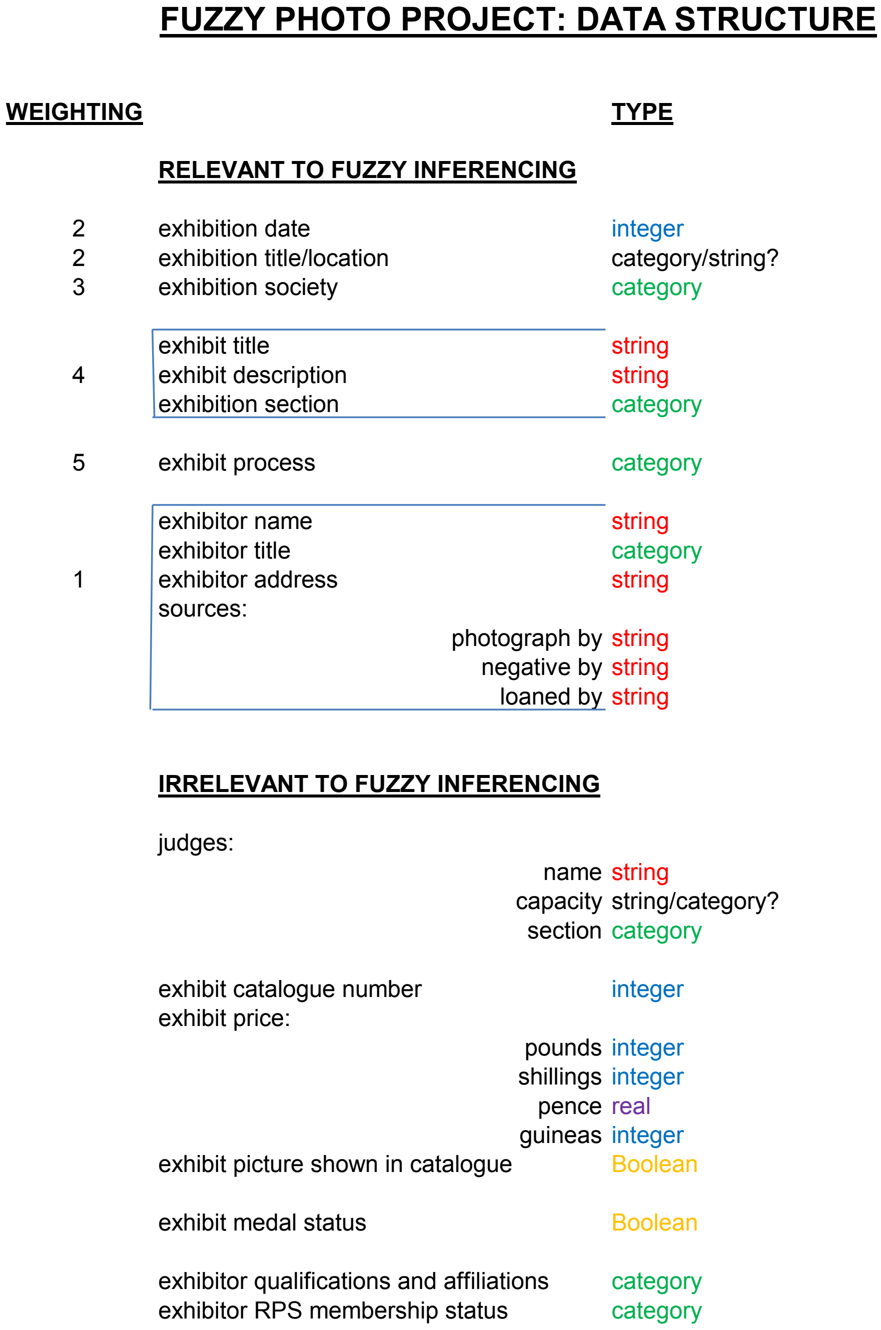
[1] In the interests of clarity the headings were changed; *DEFINITE DATA* → *RELEVANT DATA* → *RELEVANT TO FUZZY INFERENCING* and *POSSIBLE DATA* → *IRRELEVANT DATA* → *IRRELEVANT TO FUZZY INFERENCING*.



FUZZY PHOTO PROJECT: DATA STRUCTURE

	TYPE
RELEVANT DATA	
exhibition date	integer
exhibition title/location	category/string?
exhibit title	string
exhibit description	string
exhibition section	category
exhibit process	category
exhibit medal status	Boolean
exhibitor name	string
exhibitor title	category
exhibitor address	string
exhibitor qualifications and affiliations	category
exhibitor RPS membership status	category
sources:	
photograph by	string
negative by	string
loaned by	string
IRRELEVANT DATA	
exhibition society	category
judges:	
name	string
capacity	string/category?
section	category
exhibit catalogue number	integer
exhibit price:	
pounds	integer
shillings	integer
pence	real
guineas	integer
exhibit picture shown in catalogue	Boolean

Figure 3.1: The data classification used as a starting point for the interviews with Kelly Wilder, Jane Fletcher and Michael Pritchard.



FUZZY PHOTO PROJECT: DATA STRUCTURE

WEIGHTING		TYPE
	RELEVANT TO FUZZY INFERENCING	
2	exhibition date	integer
2	exhibition title/location	category/string?
3	exhibition society	category
	exhibit title	string
4	exhibit description	string
	exhibition section	category
5	exhibit process	category
	exhibitor name	string
	exhibitor title	category
1	exhibitor address	string
	sources:	
	photograph by	string
	negative by	string
	loaned by	string
	IRRELEVANT TO FUZZY INFERENCING	
	judges:	
	name	string
	capacity	string/category?
	section	category
	exhibit catalogue number	integer
	exhibit price:	
	pounds	integer
	shillings	integer
	pence	real
	guineas	integer
	exhibit picture shown in catalogue	Boolean
	exhibit medal status	Boolean
	exhibitor qualifications and affiliations	category
	exhibitor RPS membership status	category

Figure 3.2: The data as classified by Michael Pritchard.

3.2 Relevant Data Regarded as Important

3.2.1 Exhibit Data

Three fields from the DMU databases (exhibit title, exhibit description, and exhibit section) give information about the exhibit itself. In an online collection the exhibit section field will be absent. We are therefore seeking matches between 3 fields in the DMU databases and the 2 online collection fields, *image description* and *image title* (figure 3.3). Sometimes there are other searchable notes in the online collections; for the purposes of matching they will be concatenated to the description field to give *image information*

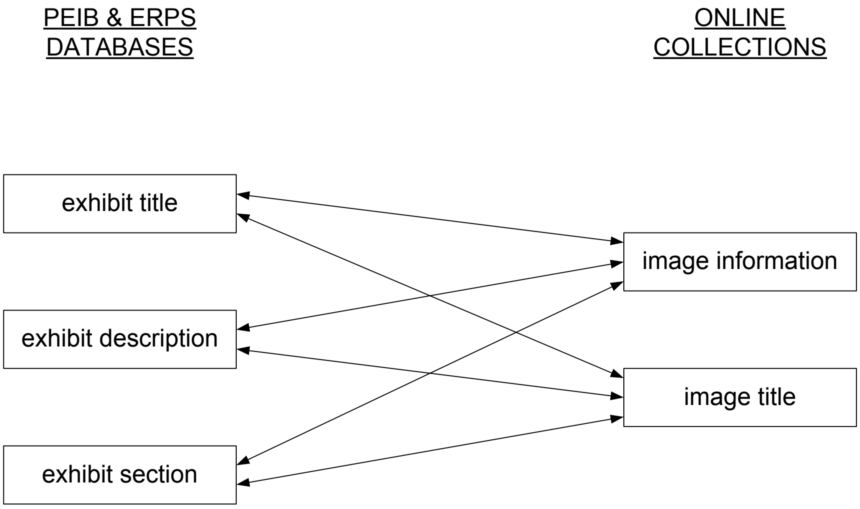


Figure 3.3: Matching exhibit data.

3.2.2 Historical Stakeholders Data

A historical stakeholder is any person who was, historically, connected to the photograph, i.e. 1. the exhibitor, 2. the person who created the negative, 3. the person who created the print from the negative, or 4. the owner of the photograph (who loaned it to the exhibitor). In the frequent cases where the four roles are performed by one individual, there is only one historical stakeholder. 'Exhibitor details' is an amalgamation of exhibitor name, title, address, qualifications and affiliations, and RPS membership status. On the online collection side, we would expect there to be one field, *photographer name*. In this case we a looking for matches between 4 fields in the DMU databases and 1 in an online collection (figure 3.4).

3.2.3 Processes Data

In the DMU research databases, process data is a category. In an online database, process data, if it exists at all, will be a substring (section B.1) of the description field.

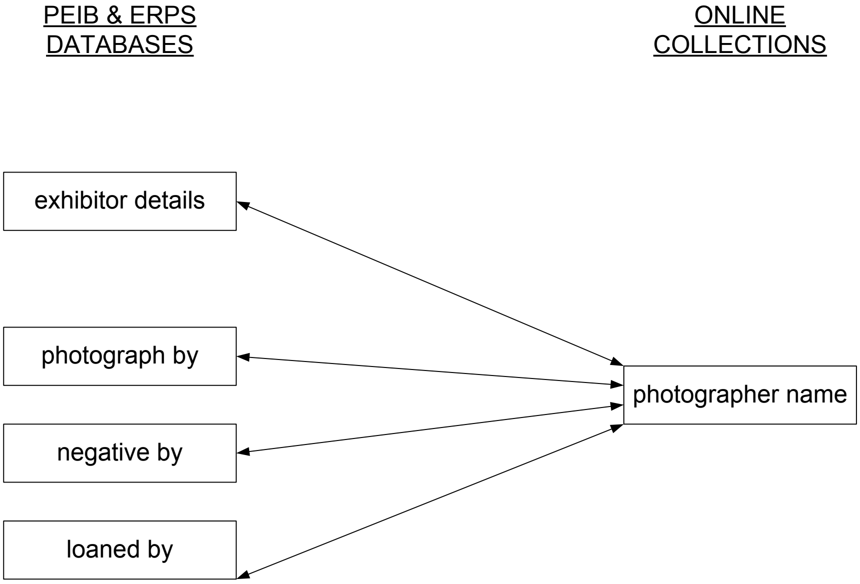


Figure 3.4: Matching historical stakeholders data.

3.2.4 Dates Data

In the DMU databases, the *date* (in years) is the date of the catalogue. In an online database, date data, as with processes data, will be a substring of the description field if it exists at all. An additional complication is that in the online records, dates may refer to any significant event connected to the photograph.

A Fuzzy Approach? We are now in a position to answer the first subsidiary question arising out of the research question. The string data (exhibit data and historical stakeholder), though not entirely reliable, may be matched solely through string matching. However given the uncertainty in the processes (categorical) and dates (numerical) data, a fuzzy approach to inferencing is recommended (subsection A.3).

3.3 Matching Strings

Before strings are ready to be matched, some pre-processing is required. This is so that

1. 1-1 string matching may be performed. This is essential.
2. Characters which will not help in matching are excluded to make the string matching algorithms run faster.
3. The matching algorithms do not need to concern themselves with distinguishing between lower case and upper case letters, and will therefore run faster.

3.3.1 Pre-Processing Exhibit Data

Algorithm 1 Pre-Processing Exhibit Data

```
Capitalise all text in the DMU and online databases.
In the DMU and online databases, extract all punctuation apart from hyphens. Leave spaces and numbers.
In the DMU databases, concatenate exhibit title, exhibit description, and exhibit section to form string $DMUphoto$, leaving a space between each concatenated field.
In the online database, concatenate image description with any searchable notes or reports to form string image information, leaving a space between each concatenated field.
In the online database, concatenate image information and image title to form string $OnlinePhoto$, leaving a space between the concatenated fields.
```

The *exhibition section* category is treated as a string.

3.3.2 Pre-Processing Historical Stakeholder Data

Algorithm 2 Pre-Processing Historical Stakeholder Data

```
Capitalise all text in the DMU and online databases.
In the DMU and online databases, extract all punctuation apart from hyphens. Leave spaces and numbers.
In the DMU databases, concatenate exhibitor name, title, address, qualifications and affiliations, and RPS membership status to create exhibitor details, leaving a space between each concatenated field.
Concatenate exhibitor details, ‘photographed by’, ‘negative by’, and ‘loaned by’ to form string $DMUperson$, leaving a space between each concatenated field.
Designate the online photographer name field $OnlinePhotographer$.
```

The *exhibitor title*, *exhibitor qualifications and affiliations* and *exhibitor RPS membership status* categories are treated as strings.

3.3.3 String Matching Strategies

After pre-processing, the problem is reduced to matching two pairs of strings (figures 3.5 and 3.6).

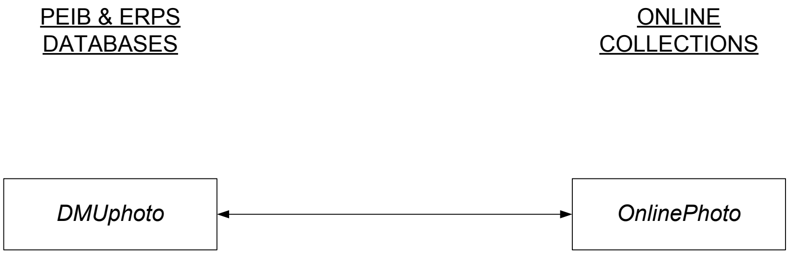


Figure 3.5: Matching pre-processed exhibit data.

The two available strategies for string matching are *exact string matching* (subsection B.2.1) and *approximate string matching* (subsection B.2.2). Exact string matching is relatively efficient, but will not

PEIB & ERPS
DATABASES
ONLINE
COLLECTIONS
DMUperson
OnlinePhotographer

Figure 3.6: Matching pre-processed historical stakeholders data.

Algorithm 3 String Matching Algorithm

```
Break the DMU database string into words (patterns), using the spaces as separators.
Taking each DMU database pattern in turn, match it with the relevant online string (OnlinePhoto or OnlinePhotographer) using exact string matching.
Taking each DMU database pattern in turn, match it with the relevant online string (OnlinePhoto or OnlinePhotographer) using approximate string matching.
```

match misspelt words. Approximate string matching will pick up misspelt words, but is not suitable for matching words of length less than 5, which would be an issue for some of the titles, some qualifications and affiliations, and for very short names. As there are many short words, as well as numerous spelling inconsistencies, a combination of the two approaches is to be recommended. The space separators will ensure that short words are not treated as parts of longer words.

For approximate string matching, a choice has to be made about the maximum error allowed (k). To allow for matching words such as 'Delamotte', which may be respelt as three words with two spaces, a value of $k = 2$ is optimal.

3.3.4 The Matching Metric

The matching metric (M) is a measure of how well the strings match. It is a number between 0 and 100, i.e. a percentage. 0 signifies that the two strings have nothing in common. $M = 100$ when two strings match exactly[2]. We suggest two versions of the matching metric, the Simple Matching Metric and the Fractional Matching Metric.

Algorithm 4 Simple Matching Metric

```
Choose $k$, the maximum error allowed.
Count $s$, the number of words in the DMU string, that are matched within the allowable error.
Count $t$, the total number of words in the DMU string.
$M = \frac{s}{t} \times 100$.
```

[2] If two strings contain exactly the same words, but in a different order, M will still be 100.

Algorithm 5 Fractional Matching Metric

```
Choose $k$, the maximum error allowed.
For each word matched within the allowable error, calculate the error level (subsection B.2.2).
Add up the error levels of each matched word to give $s$.
Count $t$, the total number of words in the DMU string.
$M = \frac{s}{t} \times 100$.
```

3.4 Building a Fuzzy Inferencing System

3.4.1 Exhibit and Historical Stakeholders Data

We saw in subsection 2.3 that the *exhibit* and *historical stakeholders* data were less than optimal. The purpose of the matching metric is to automatically make string matching fuzzy (without the need for expert input). There are a number of ways of accomplishing this.

3.4.2 Processes

The processes data is incomplete and extremely unreliable in the DMU research databases (subsection 3.2.3), and probably equally so in the online collection databases. Therefore fuzzy logic is ideally suited to matching processes data. In the online collections, the process data, if it exists, is embedded in the description field and will need to be extracted.

It will be necessary to re-interview the experts in relation to the processes data. There are two slightly different fuzzy logic approaches possible:

1. To create type-1 fuzzy sets, each expert would be asked to rate the reliability of the data for each process on a scale from 1 to 10.

2. To create type-2 fuzzy sets, each expert would be asked to rate the reliability of the data for each process on a natural language based scale, starting at *totally unreliable*, ending at *totally reliable*, and passing through options such as *quite unreliable* and *quite reliable*.

A fuzzy technique for resolving differences of opinion among experts is [3].

The six processes in the ERPS drop-down menu are umbrella terms for a larger array of processes. In the PEIB database the processes have not been categorised, and this may or may not be the case with the online collections databases. To get round this it may be advisable to match using the umbrella terms *and* the original process names.

3.4.3 Dates

In the DMU research databases, the dates data is complete and reliable. However in the online databases the dates data is incomplete and there are many instances of estimates (subsection 3.2.4). In the online catalogues dates need to be extracted from the *description* field. As with processes, fuzzy logic is ideal for matching the dates data.

It might also be useful for the experts to re-visit the question of how close two dates would have to be for them to be significant in the matching process. There are two considerations:

Dates Excluding a Match If the date in the DMU database precedes the date (of creation) in the online catalogue, then the two photographs cannot match. However, this presupposes that the online date is accurate, when in fact it may have been an estimate.

Retrospective Exhibitions If dates from retrospective exhibitions are included in the online data, this can give the impression that two photographs do not match, when in fact they do.

3.4.4 Type-2 versus Type-1

A type-2 FIS is to be preferred over a type-1 FIS. This is because the type-2 fuzzy set has an additional (third) dimension with which to specifically model the uncertainties associated with data [25]. The uncertainties and omissions within the photographic records are challenging; the sophisticated modelling of a type-2 fuzzy set is the best response to this challenge.

3.4.5 Confidence Level

The *confidence level* of suggested matches would be a vital piece of information. Let us define the confidence level of a potential match to be a number between 0 and 1, with 1 representing total confidence. As the confidence level increases, the uncertainty decreases. To work out the confidence level begin by calculating the uncertainty associated with the aggregated fuzzy set (subsection A.5)[3], and then subtract it from 1 to give the confidence level. As a formula,

$$C = 1 - U$$

where C is the confidence level and U is the uncertainty. This will give a number between 0 and 1; to convert it to a percentage, simply multiply by 100.

[3] If necessary, normalise the result to between 0 and 1 i.e. adjust the range of possible uncertainty values to start at 0 and end at 1, then fit the actual uncertainty value proportionately within this range.

Chapter 4

Which Metadata Schema?

4.1 Metadata

If matching information online from two different databases, it is advantageous if they have both been created within the constraints of the same metadata schema. The DMU *Photohistory* databases employ Dublin Core metadata (DC) to describe their contents. This is to make them more easily discoverable by external search engines and to facilitate data aggregation between these and other resources such as the Arts and Humanities Data Service (AHDS). DC has the advantages that it is easy to use and widely adopted. An alternative approach could be to use CIDOC-CRM. CRM stands for 'Conceptual Reference Model'; it was developed by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). CIDOC-CRM has the potential to add semantic meaning to the metadata mark-up, thus enhancing the potential for identifying meaningful links between different resources. The semantic capability of CIDOC-CRM derives from the fact that it is ontology based. *Ontology* is the branch of philosophy concerned with *what there is*, i.e. what exists. In computer science, ontologies are developed for specific domains of knowledge such as anatomy, pathology, architecture etc.. A database may be created with or without reference to an ontology. Within an ontology, it is possible to describe the domain as well as reason about the subject matter of the domain. Ontologies are crafted by panels of domain experts, who consider in depth the requirements of the community that is to use the ontology. They are reviewed and refined in the light of feedback from the ontology users.

4.2 Dublin Core

The Dublin Core Metadata Terms are designed for librarians, i.e. with books and other written texts in mind. Dublin Core is an unstructured list of elements (subsection C.1.1).

4.2.1 How the Photographic Records Fit into Dublin Core

The current DMU research databases attempt to follow the constraints of Simple Dublin Core in assigning each of the main DMU data fields to a Dublin Core element. The assignments achieved are mostly satisfactory, but there are two major anomalies:

1. The *process* is assigned to the Dublin Core *type* element. This is according to neither the spirit nor the letter of Dublin Core, since *type* is defined as 'The nature or genre of the resource.' [17],
2. The *loaned by* data has not been assigned to any Dublin Core element. Though the *loaned by* data is in the DMU databases, it is not in the Dublin Core ontology.

The latest extension of Dublin Core, in spite of having over 50 elements, is still unable to assign *process* and *loaned by* data suitably. Lin at al. concur that Dublin Core is inadequate for cultural heritage.

> “It [Dublin Core] cannot describe causal relationships, processes or phases, such as observations or research activities that can be related to a cultural object.” [23]

4.3 CIDOC-CRM

CIDOC-CRM is specifically and successfully designed for cultural heritage applications:

> “Among various domain-specific ontologies . . . CIDOC CRM . . . is particularly well-designed for cultural heritage applications, and the only one that has become an International Standard.” [23]

A semantic net, CIDOC-CRM goes a lot further than merely listing attributes of an object:

> “It [CIDOC-CRM] is not a fusion of existing formats, but a product of expert insight and intensive interdisciplinary work. As such, it builds on a metaschema of fundamental categories and causal relationships with explanatory power.” [9]

4.3.1 How the Photographic Records Fit into CIDOC-CRM

CIDOC-CRM [5] has 90 classes covering cultural heritage in every form. Only a small subset are relevant to historical photographic data; the ten main ones are shown in figure 4.1. Various paths can be traced through the semantic net by following the arrows to superclasses (subsection C.1.2). Classes inherit from their superclasses. So for example a *Man-Made Object* (E22) **is a** *Physical Object* (E19), which **is a** *Physical Thing* (E18), and so on. In this way, information about a print, for example, is accreted from the *Man-Made Object* class and **all its superclasses**. There is an unresolved tension between *photograph as artifact* and *photograph as image*, but with CIDOC-CRM you can have it both ways. The class *E22 Man-Made Object* pertains to artifacts, whereas the class *E38 Image* clearly relates to images ([5], page 17). The print itself is covered by E22, but the image on the print belongs in E38, together with the image on its negative and the image from any other prints made from that negative.

Doerr [8] has demonstrated how Simple Dublin Core may be mapped into CIDOC-CRM. This immediately tells us that **within the CRM there are classes equivalent to all those in Dublin Core**. The current versions of the DMU research databases use Dublin Core and it is useful to know that ‘translation’ into CIDOC-CRM is possible.

Dimension, a factor that can help with photograph matching, is not an element of Dublin Core, but is a class in CIDOC-CRM (E54 Dimension).

4.3.2 Comparing Dublin Core and CIDOC-CRM

Dublin Core is not designed for historical photographic data, and does not provide a good model. CIDOC-CRM on the other hand, is designed for cultural heritage, and successfully models the data to be used in photograph matching.

Which Metadata Schema? We can now answer the second subsidiary research question. Dublin Core is inadequate for handling the photographic metadata whereas CIDOC-CRM is admirably suited to the task.

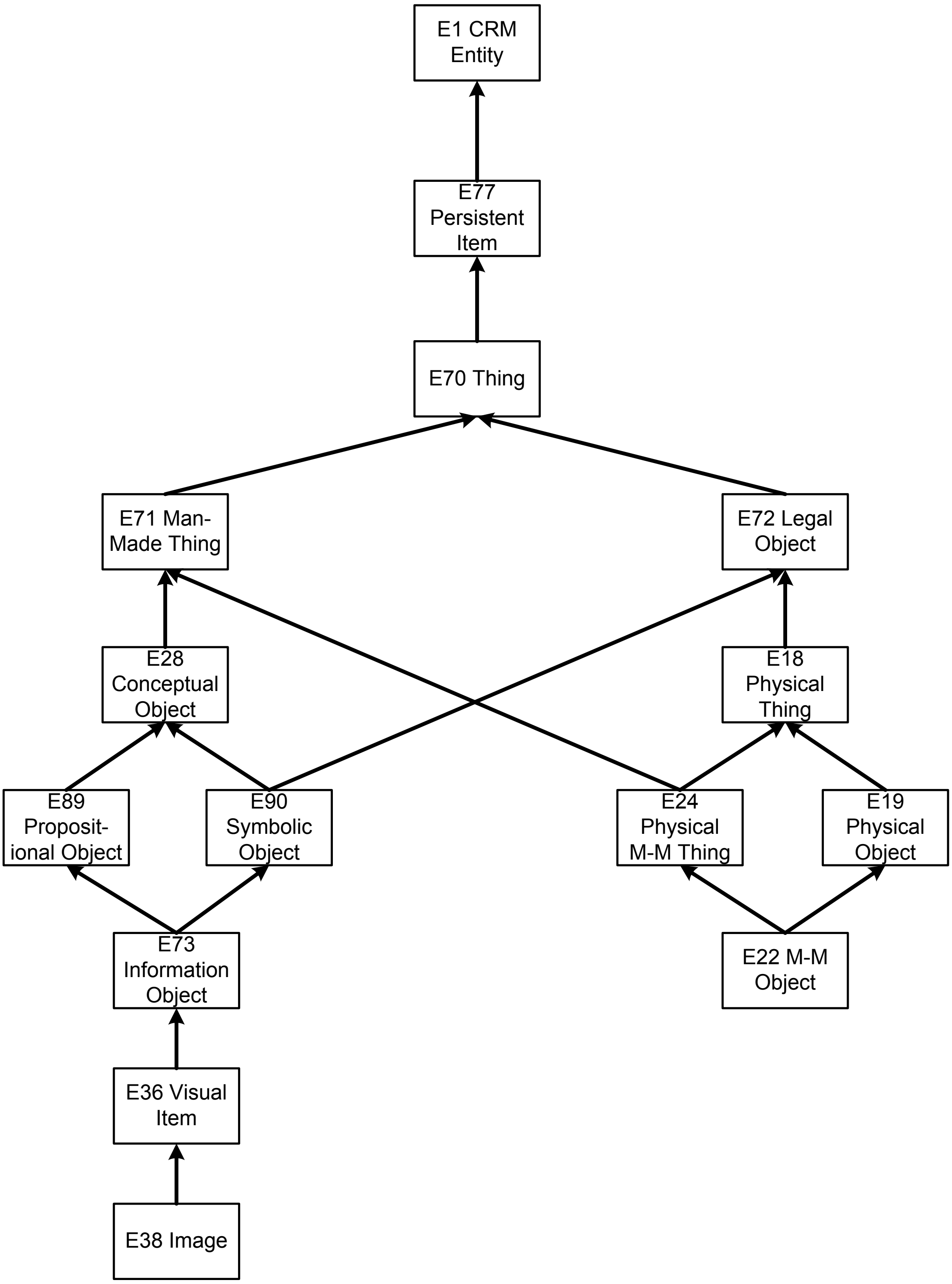


Figure 4.1: How photographs fit into CIDOC-CRM.

Chapter 5

Which Type of Database?

A *database* is simply "an organized collection of data for one or more multiple uses." [43] Relational databases are currently the most commonly used database. Information is recorded as tables, making it "readily and easily searched through" [43] by a computer and easily understood by a human being. The XML/RDF database (subsection C.3.1) is another widely used technology. Which is more appropriate for photograph matching — the relational database or the XML/RDF database? There are two issues to consider in making this decision, firstly *performance*, and secondly *compatability with fuzzy logic*.

5.1 Performance

We saw in chapter 4 that CIDOC-CRM is the ideal metadata schema for photograph matching. Of necessity CIDOC-CRM is implemented as an XML/RDF database, and this in itself is a powerful reason for preferring an XML database over a relational one. However, before the XML database can be wholeheartedly recommended, its performance has to be shown to be adequate as, "Performance is often the first critical factor when selecting a database." [30]

Performance Tests In 2006 the *Oracle Berkeley DB*, an XML database, was subjected to extensive benchmark tests to see how fast it would run under various stringent sets of conditions [30]. Performance is dependant on the operating system used, so each test was repeated with different operating systems. The first test measured single-record read times and single-record write times using Berkeley DB Data Store (DS). Table 5.1 shows the results:

Table 5.1: Measuring throughput as operations per second using the Berkeley DB Data Store (DS).

DS (ops/sec)	Linux	Solaris	Win XP	BSD	Mac OS/X
Single-record read	1,002,200	1,008,580	1,000,000	1,108,920	524,360
Single-record write	766,034	550,748	447,628	614,116	317,141

These times are undoubtedly fast times. Other tests were run under differing constraints; they still gave impressive results [30]. Though not providing a direct comparison with the relational database, these tests show the XML/RDF alternative to run sufficiently speedily to be practicable.

5.2 Compatability with Fuzzy Logic

5.2.1 Fuzzy RDF

Work is progressing in the theory and application of Fuzzy RDF (subsection C.2.3). RDF consists of triples. The obvious way of implementing Type-1 Fuzzy RDF is to associated a membership grade with an RDF triple [36]. Type-2 Fuzzy RDF could be implemented as a triple associated with a type-1 fuzzy set. Li at al. [21] have devised a more complex structure for Type-2 Fuzzy RDF, with good experimental results. There is clearly no obstacle to the development of Fuzzy RDF.

Which Type of Database System for Matching? We can now answer the third subsidiary research question. The XML/RDF database is recommended for storing the metadata to be matched because 1. it runs sufficiently fast, and 2. RDF is extendible to Fuzzy RDF.

5.3 Storing the Results of the Matching Process

The matched results gained by running the FPM may be recorded in a lengthy yet simple table (as shown in figure 5.2). The first three columns taken together specify a catalogue entry. Thereafter there is a column for each online collection. Each online collection column shows an ordered pair of numbers, representing a potential match. The first number of the ordered pair is the online collection reference number, the second the confidence level of the match.

A photographic historian searching this database will be able to specify a catalogue entry (using *DMU Database*, *Year* and *Catalogue Number*) and a *minimum confidence level*. A list will then be displayed showing potential matches corresponding to the parameters.

Table 5.2: Database of output of the Fuzzy Photo Matcher. (The two tabulated examples are invented.)

DMU Database	**Year**	**Catalogue Number**	**Online Collection 1**	**Online Collection 2**	**Online Collection 3**	**...** **...**
⋮	⋮	⋮	⋮	⋮	⋮	...
PEIB	1845	22	(55,0.61)	(72,0.30)	(111,0.75)	...
⋮	⋮	⋮	⋮	⋮	⋮	...
ERPS	1905	56	(16,0.56)	(18,0.88)	(109,0.34)	...
⋮	⋮	⋮	⋮	⋮	⋮	...

Which Type of Database System for Storing Matches? We can now answer the fourth and last subsidiary research question. A relational database system consisting of a single table is ample for storing the Fuzzy Photo Matcher data.

Chapter 6

Conclusion

6.1 The Research Questions Answered

Four questions were posed in subsection 1.1.1, and answered within this report.

1. In chapter 3 it was argued that fuzzy logic is the best strategy for matching DMU photographic research database data with that of online collections, and that on its own string matching is insufficient.

2. In chapter 4 CIDOC-CRM was shown to be the most appropriate metadata schema for the development of historic photographic databases.

3. If CIDOC-CRM is to be implemented, it has to be as an XML/RDF database. Chapter 5 provided reassurance that such as implementation is feasible and compatible with fuzzy logic.

4. In chapter 5 a relational database was recommended as appropriate for storing the output of the FPM.

6.2 Recommended Future Directions of Research

Embedded Data An exhibit title may refer to a place or person (section 2.5. For the researcher of photographic history, there may be useful links with online databases such as the *Getty Thesaurus of Geographic Names Online* [10], or *Debretts* [7] (historical notable people).

Matching Negatives There are thousands of historical negatives in collections. The FPM could be extended to match negatives with their prints.

History of Photography Linked to History of Art Since the invention of photography, certain individuals have been regarded as both artists and photographers. They either see photography as a form of art, or take photographs from which to paint portraits. Furthermore, prints are frequently made from famous paintings. It is conceivable that the FPM technology may be used to linking the history of art with the history of photography.

6.3 Unintended Benefits of the FPM

The FPM, if and when implemented, should achieve unforeseen results:

- Matching photographs in two different online collections if they both match a DMU database photo.
- Matching photographs from two PEIB or ERPS exhibitions.

References

[1] Max Black. Vagueness. In *Philosophy of Science, pp. 427-455*. 1937.

[2] William W. L. Cheung, Tony J. Pitcher, and Daniel Pauly. A Fuzzy Logic Expert System to Estimate Intrinsic Extinction Vulnerabilities of Marine Fishes to Fishing. *Biological Conservation*, 124:97–111, 2005.

[3] Francisco Chiclana, Francisco Herrera, and Enrique Herrera Viedma. Integrating Three Representation Models in Fuzzy Multipurpose Decision Making Based on Fuzzy Preference Relations. *Fuzzy Sets and Systems*, 97(1):33–48, 1998.

[4] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, third edition*. The MIT Press, Cambridge, Massachusetts, USA, 2009.

[5] Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff. *Definition of the CIDOC Conceptual Reference Model*. ICOM/CIDOC CRM Special Interest Group, 2010. version 5.0.2.

[6] John Davies, Dieter Fensel, and Frank van Harmelen. *Towards the Semantic Web: Ontology Driven Knowledge Management*. Wiley, 2003.

[7] Debrett's. *Debrett's*. http://www.debretts.com/home.aspx [Online; accessed 30-July-2010].

[8] Martin Doerr. Mapping of the Dublin Core Metadata Set to the CIDOC CRM. Technical report, Institute of Computer Science, Foundation for Research and Technology – Hellas Science and Technology Park of Crete, P. O. Box 1385, GR 711 10, Heraklion, Crete, Greece, July 2000.

[9] Martin Doerr, Nick Crofts, and Maria Theodoridou. Metadata and the CIDOC-CRM — a Solution for Semantic Interoperability, 2000. Presented at CIDOC 2000, Ottowa. www.chin.gc.ca/Resources/Cidoc/French/Presentations/mdoerr.ppt [Online; accessed 30-July-2010].

[10] The Getty. *The Getty Thesaurus of Geographic Names Online*. http://www.getty.edu/research/conducting_research/vocabularies/tgn/ [Online; accessed 30-July-2010].

[11] Sarah Greenfield. *Uncertainty, Imprecision and Vagueness*. MSc Thesis, De Montfort University, UK, 2005.

[12] Sarah Greenfield and Robert I. John. *The Uncertainty Associated with a Type-2 Fuzzy Set*, volume 243 of *Studies in Fuzziness and Soft Computing*, chapter 23, pages 471–483. Springer-Verlag, 2009.

[13] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, New York, NY, USA, 1997.

[14] Petr Hajek. Fuzzy Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/fall2002/entries/logic-fuzzy .

[15] Johan Hjelm. *Creating the Semantic Web with RDF: Professional Developer's Guide*. Wiley, USA, 2001.

[16] George Eastman House. *George Eastman House: International Museum of Photography and Film*. http://www.eastmanhouse.org/inc/collections/photography.php [Online; accessed 9-July-2010].

[17] Dublin Core Metadata Initiative. DCMI Metadata Terms, 2010. http://dublincore.org/documents/dcmi-terms/ [Online; accessed 26-July-2010].

[18] Jyh-Shing Roger Jang. ANFIS: Adaptive–Network–Base Fuzzy Inference System. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3):665–685, 1993.

[19] George J. Klir and Tina A. Folger. *Fuzzy Sets, Uncertainty, and Information*. Prentice-Hall International, 1992.

[20] Vladimir Levenshtein. Binary Codes Capable of Correctiong Deletions, Insertions, and Reversals. *Sov. Phys. Dokl.*, 10(8):707–710, 1966. Original in Russian in *Dokl. Akad. Nauk SSSR* 163, 4, 845–848, 1965.

[21] Ruixuan Li, Xiaolin Sun, Zhengding Lu, Kunmei Wen, and Yuhua Li. *Towards a Type-2 Fuzzy Description Logic for Semantic Search Engine* , volume 4505 of *Lecture Notes in Computer Science*, pages 805–812. Springer Berlin/Heidelberg, 2007.

[22] Jesse Liberty and Mike Kraley. *XML Web Documents from Scratch*. Que Corporation, USA, 2000. ISBN: 0-7897-2316-6.

[23] Chia-Hung Lin, Jen-Shin Hong, and Martin Doerr. Issues in an Inference Platform for Generating Deductive Knowledge: a Case Study in Cultural Heritage Digital Libraries using the CIDOC-CRM. *International Journal on Digital Libraries*, 8:115–132, 2008.

[24] Aimilia Magkanaraki, Sofia Alexaki, Vassilis Christophides, and Dimitris Plexousakis. Benchmarking RDF Schemas for the Semantic Web. In *In Proc. of the International Semantic Web Conference (ISWC)*, pages 132–146, 2002.

[25] J. M. Mendel and Robert I. John. Type-2 Fuzzy Sets Made Simple. *IEEE Transactions on Fuzzy Systems*, 10(2):117–127, 2002.

[26] Jerry M. Mendel. *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions*. Prentice-Hall PTR, 2001.

[27] UKOLN Michael Day. Metadata in a Nutshell, 2010. http://www.ukoln.ac.uk/metadata/publications/nutshell/ [Online; accessed 4-September-2010].

[28] Gonzalo Navarro. A Guided Tour to Approximate String Matching. *ACM Computing Surveys*, 33(1):31–88, March 2001.

[29] NISO. Metadata, 2010. http://framework.niso.org/node/24 [Online; accessed 4-September-2010].

[30] Oracle. Oracle Berkeley DB: Performance Metrics and Benchmarks, *An Oracle White Paper.*, August 2006. http://www.oracle.com/technology/products/berkeley-db/pdf/berkeley-db-perf.pdf [Online; accessed 19-July-2010].

[31] Erik T. Ray. *Learning XML*. O'Reilly and Associates, USA, 2001. Print ISBN: 978-0-596-00046-2; ISBN 10: 0-596-00046-4.

[32] Jack Rusher. Triple Store, 2010. http://www.w3.org/2001/sw/Europe/events/20031113-storage/positions/rusher.html [Online; accessed 2-September-2010].

[33] Moti Schneider, Horst Bunke, and Abraham Kandel. Using Fuzzy Logic to Match Strings in Documents. *International Journal of Intelligent Systems*, 16:609–619, 2001.

[34] startvbdotnet.com. Extensible Markup Language (XML), 2010. http://www.startvbdotnet.com/xml/ [Online; accessed 1-June-2010].

[35] Umberto Straccia. A Minimal Deductive System for General Fuzzy RDF. In *Proceedings of 3rd International Conference on Web Reasoning and Rule Systems*, July 2009.

[36] Octavian Udrea, Diego Reforgiato Recupero, and V. S. Subrahmanian. Annotated RDF. *ACM Transactions on Computational Logic*, V(N):1–40, 2008.

[37] UKOLN. Metadata, 2010. http://www.ukoln.ac.uk/metadata/ [Online; accessed 4-September-2010].

[38] Veronika Vaneková, Ján Bella, Peter Gurský, and Tomás Horváth. Fuzzy RDF in the Semantic Web: Deduction and Induction. In *Proceedings of 6th Workshop on Data Analysis*, Abaujszanto, Hungary, June 2005.

[39] W3C. Resource Description Framework (RDF) Model and Syntax Specification, 2010. http://www.w3.org/TR/PR-rdf-syntax/ [Online; accessed 2-September-2010].

[40] w3schools. RDF Schema (RDFS), 2010. http://www.w3schools.com/rdf/rdf_schema.asp [Online; accessed 22-September-2010].

[41] w3schools. XML Schema Tutorial, 2010. http://www.w3schools.com/schema/default.asp [Online; accessed 1-June-2010].

[42] Wikipedia. Approximate string matching — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/wiki/Approximate_string_matching [Online; accessed 21-June-2010].

[43] Wikipedia. Database — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/wiki/Database [Online; accessed 10-June-2010].

[44] Wikipedia. Dublin Core — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/wiki/Dublin_Core [Online; accessed 19-June-2010].

[45] Wikipedia. Markup language — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/wiki/Markup_language [Online; accessed 1-June-2010].

[46] Wikipedia. Neuro-fuzzy — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/wiki/Neuro-fuzzy [Online; accessed 6-July-2010].

[47] Wikipedia. Ontology (information science) — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/w/index.php?title=Ontology_(information_science)&oldid=354929096 [Online; accessed 10-April-2010].

[48] Wikipedia. RDF Schema — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/w/index.php?title=RDF_Schema&oldid=350843550 [Online; accessed 10-April-2010].

[49] Wikipedia. Resource Description Framework — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/wiki/Resource_Description_Framework [Online; accessed 22-September-2010].

[50] Wikipedia. Semantic Web — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/w/index.php?title=Semantic_Web&oldid=354927878 [Online; accessed 9-April-2010].

[51] Wikipedia. Serialization — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/wiki/Serialization#cite_note-0 [Online; accessed 2-June-2010].

[52] Wikipedia. SPARQL — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/w/index.php?title=SPARQL&oldid=350844116 [Online; accessed 10-April-2010].

[53] Wikipedia. String (computer science) — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/wiki/String_(computer_science) [Online; accessed 14-June-2010].

[54] Wikipedia. String searching algorithm — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/wiki/String_searching_algorithm [Online; accessed 15-June-2010].

[55] Wikipedia. Triplestore — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/wiki/Triplestore [Online; accessed 2-September-2010].

[56] Wikipedia. Typographical error — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/wiki/Typographical_error [Online; accessed 5-July-2010].

[57] Wikipedia. XML — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/wiki/XML [Online; accessed 23-July-2010].

[58] Wikipedia. XML database — Wikipedia, The Free Encyclopedia, 2010. http://en.wikipedia.org/wiki/XML_database [Online; accessed 10-June-2010].

[59] Dongrui Wu and Jerry M. Mendel. Uncertainty Measures for Interval Type-2 Fuzzy Sets. *Information Sciences*, 177:5378–5393, 2007.

[60] Lotfi Zadeh. Fuzzy Sets. In *Information and Control 8, pp. 338-353*. 1965.

Appendix A

Fuzzy Logic

A.1 Fuzzy Set Theory

Fuzzy set theory was originated by Lotfi Zadeh [60] in the 1960s. As far back as 1937, however, Max Black [1] had proposed a similar, though not identical idea, using the terminology of *vague sets*. At that time his work remained obscure. For more on the motivation and philosophy underpinning the concept of fuzziness, see [11].

Set theory is concerned with whether an object satisfies a specific description. Fuzzy set theory is concerned with the extent to which an object satisfies an inexact description: a fuzzy set is simply a set that does not have sharp boundaries, unlike the conventional crisp set. Someone might know all the relevant facts, yet still be equivocal about whether a description applies to a specific object. The classic example would be that of a man of $5'10''$, the question being whether he would be described as tall. Though such a man would be taller than average, most people would probably be hesitant to describe him as tall, though many might describe him as *quite* tall. Suppose the man in question is called Peter. Most people would be uncomfortable with the idea of 'Peter is tall.' being an unequivocally true statement. However fuzzy set theory allows for degree of truth. Truth-values form a continuum on a scale from 0 to 1, with 0 representing *false*, and 1 representing *true*. The fuzzy viewpoint would permit one to say

$$\mathit{Tall}(\mathit{Peter}) = 0.9,$$

thus encapsulating the notion of Peter being on the tall side, but not being remarkably tall. In contrast, the statement

$$\mathit{Tall}(\mathit{Paul}) = 0.99,$$

would indicate that Paul was particularly tall - so tall in fact that virtually nobody would be uncomfortable with his being described as tall.

A.1.1 Membership Functions

Every fuzzy set is associated with a *membership function* – it is through its membership function that a fuzzy set is defined. The membership function maps each element of the domain onto its *degree of membership*, i.e. its truth-value. Figures A.1 to A.4 show graphical representations of membership functions.

A.1.2 Type-2 Fuzzy Sets

The fuzzy sets described so far are the 'ordinary' or type-1 variety, whose membership functions take crisp values. The derivation of a type-1 membership function tends to be a subjective process, involving

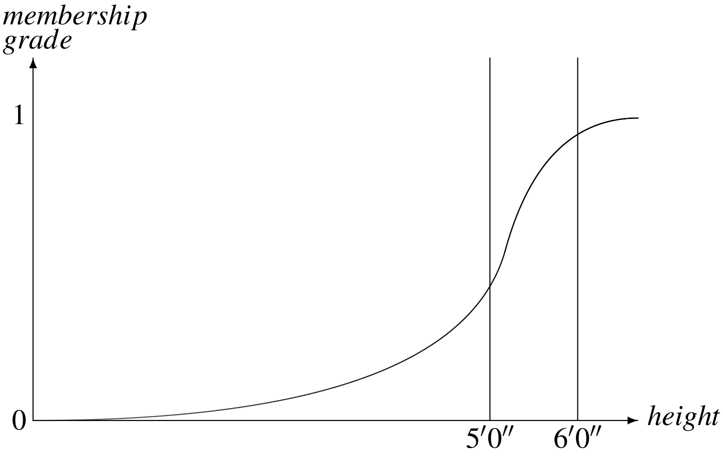


Figure A.1: The membership function for *tall*.

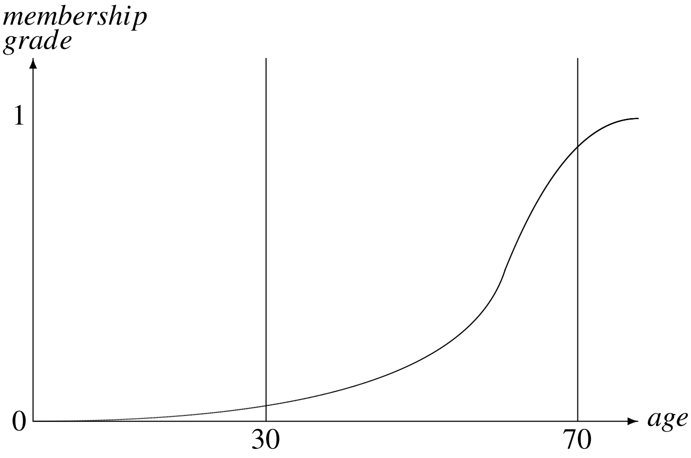


Figure A.2: The membership function for *old*.

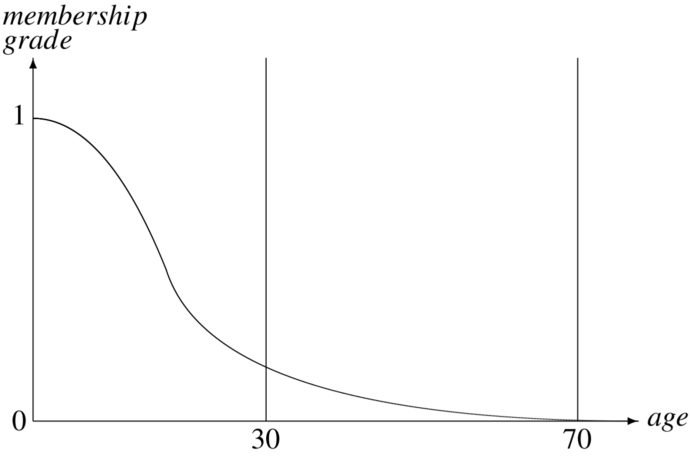


Figure A.3: The membership function for *young*.

considerable guesswork and intuition. Moreover, using real numbers (possibly expressed to several decimal places) to represent degrees of membership of fuzzy sets seems rather counterintuitive. Klir and Folger [19], page 12 comment:

> "... it may seem problematical, if not paradoxical, that a representation of fuzziness is made using membership grades that are themselves precise real numbers. Although this does not pose a serious problem for many applications, it is nevertheless possible to extend the concept of the fuzzy set to allow the distinction between grades of membership to become blurred. Sets described in this way are known as *type 2 fuzzy sets*."

Thus type-2 fuzzy sets have elements whose membership grades are themselves fuzzy sets (of type-1). It follows that the graph of a type-2 fuzzy set is 3-dimensional. Figure A.11 show a type-2 fuzzy set (from a Matlab^{TM} application), together with its Footprint of Uncertainty (FOU), which is the projection of the type-2 set onto the $x-y$ plane. Figures A.5 and A.6 depict the same FOU, with figure A.6 indicating two vertical slices at x_1 and x_2.

Secondary Membership Functions

Operations on Fuzzy Sets The operations that may be performed on crisp sets are union, intersection, and complement. Operations on fuzzy sets are developed out of the corresponding operations on crisp sets, but are more complex, involving for union and intersection the use of the *maximum* and *minimum* functions.

A.2 Fuzzy Logic

The relationship between conventional set theory and conventional logic is such that the set-theoretic statement

$Leicester \in \{cities\}$,

may be translated into the proposition

Leicester is a city.

There is a similar relationship between fuzzy sets and fuzzy propositions. Fuzzy logic is the calculus of fuzzy propositions. Under the usual formulation the interval of numbers between 0 and 1 inclusive represents the degrees of truth of propositions, with 1 denoting absolute truth, and 0 absolute falsity. Thus fuzzy logic is a form of many-valued logic.

A.2.1 Historical and Philosophical Aspects

Hajek distinguishes two strands in fuzzy logic:

> "*Fuzzy logic in the broad sense* (older, better known, heavily applied but not asking deep logical questions) serves mainly as apparatus for fuzzy control, analysis of vagueness in natural language and several other application domains. It is one of the techniques of *soft-computing*, i.e. computational methods tolerant to suboptimality and imprecision (vagueness) and giving quick, simple and *sufficiently good* solutions. ... *Fuzzy logic in the narrow sense* is symbolic logic with a comparative notion of truth developed fully in the spirit of classical logic (syntax, semantics, axiomatization, truth-preserving deduction, completeness, etc.; both propositional and predicate logic). It is a branch of *many-valued logic* based

> on the paradigm of *inference under vagueness*. This fuzzy logic is a relatively young discipline, both serving as a foundation for the fuzzy logic in a broad sense and of independent logical interest, since it turns out that strictly logical investigation of this kind of logical calculi can go rather far." [14]

The foundational assumption of classical, two-valued, logic, that every proposition is either true or false, has been subject to controversy going back to Aristotle. Klir and Folger provide the background to three-valued logic:

> "In his treatise *On Interpretation*, Aristotle discusses the problematic truth status of matters that are future-contingent. Propositions about future events, he maintains, are neither actually true not actually false but are potentially either; hence, their truth-value is undetermined, at least prior to the event.
>
> It is now well understood that propositions whose truth status is problematic are not restricted to future events. As a consequence of the Heisenberg principle of uncertainty, for example, it is known that truth-values of certain propositions in quantum mechanics are inherently indeterminate due to fundamental limitations of measurement. In order to deal with such propositions, we must relax the true-false dichotomy of classical two-valued logic by allowing a third truth-value, which may be called *indeterminate*.
>
> The classical two-valued logic can be extended into *three-valued logic* in various ways. Several three-valued logics, each with its own rationale, are now well established." [19], page 27

A.2.2 Fuzzy Inferencing Systems

A Fuzzy Inference System (FIS) is a computerized inference system which applies fuzzy logic to common-sense rules. Five processes comprise the fuzzy inference process, namely fuzzification, logical operation, implication, aggregation, and defuzzification.

A.3 The Fuzzy Inferencing Process

A Fuzzy Inferencing System works by applying fuzzy logic operators to common-sense linguistic rules. It starts with a crisp number[1], and passes through three stages: fuzzification, inferencing, and finally defuzzification:

1. "Fuzzification is a process that determines the degree of membership to the fuzzy set based on the fuzzy membership function." [2]
2. Inferencing is the main stage of calculation, the output which is a fuzzy set.
3. During the defuzzification stage this fuzzy set is converted into another crisp number, the 'answer' to the problem presented to the FIS. In a type-2 FIS, the defuzzification stage consists of two parts, type-reduction and defuzzification proper.

Figure A.12 provides a representation of a type-2 FIS.[2] The differences between this and a type-1 FIS are that 1. type-2 fuzzy sets are used, and 2. there is an extra stage of type-reduction.

[1] A crisp number is a normal number. The term 'crisp' contrasts with 'fuzzy'.

[2] An FLS (or Fuzzy Logic System) is another term for an FIS.

A.4 Learning the Rules of an FIS

In *expert driven learning*, (*knowledge elicitation*), subject experts explain, in a knowledge elicitation interview, the rules they employ to reach their conclusions.

In *data driven learning* the content of the rules is determined by learning from historical data using an automated technique such as neural networks ([46], [18]). The historical data consists of previous instances of inferences performed by experts. So both forms of rules learning are ultimately dependant on experts.

Membership Function Derivation As with rules, derivation of membership functions may be either expert driven or data driven.

A.5 Calculating the Uncertainty Associated with a Fuzzy Set

The aggregated set (subsection A.2.2) is an indicator of the uncertainty associated with the fuzzy inferencing result.

Type-1 Sets Wu and Mendel's paper "Uncertainty Measures for Interval Type-2 Fuzzy Sets" [59] lists the various uncertainty measures available for type-1 sets (page 5379).

Type-2 Sets Greenfield and John [12] describe and justify the volume measure of uncertainty for type-2 fuzzy sets.

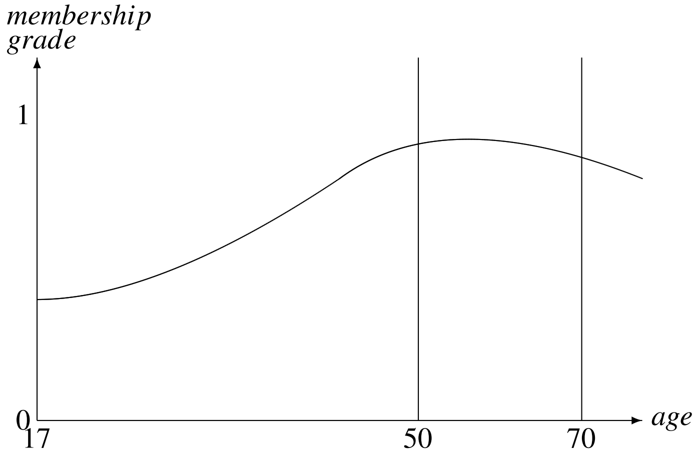


Figure A.4: The membership function for *good_driver*.

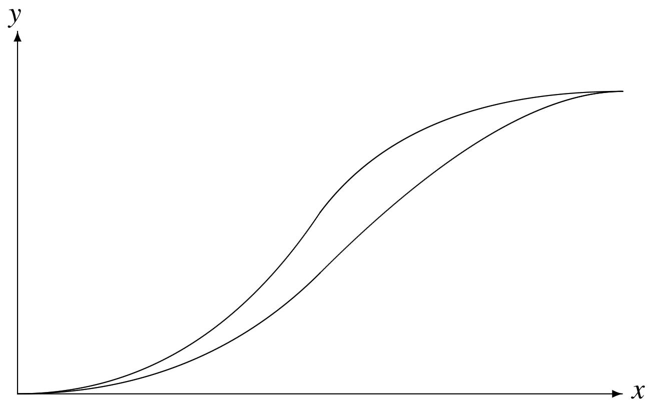


Figure A.5: FOU of a type-2 fuzzy set.

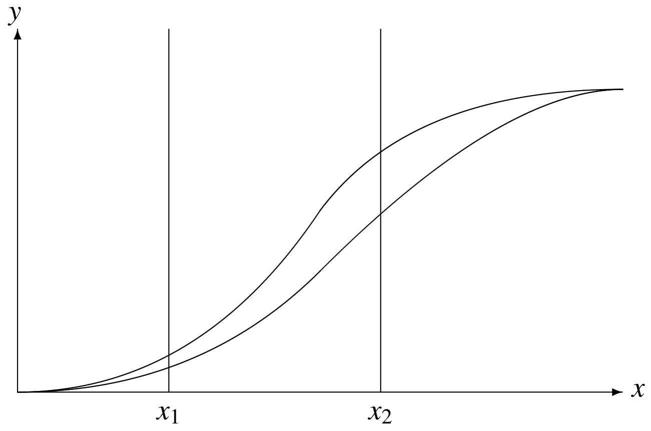


Figure A.6: FOU from figure A.5, with two vertical slices at x_1 and x_2.

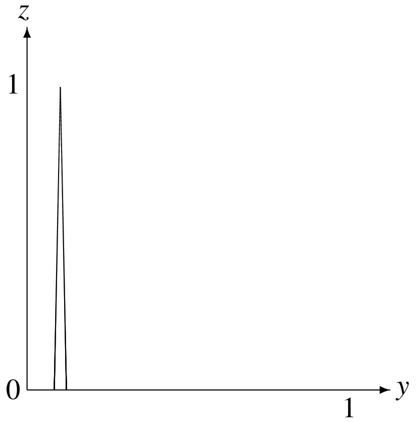


Figure A.7: Triangular secondary membership function of the vertical slice x_1 (figure A.6).

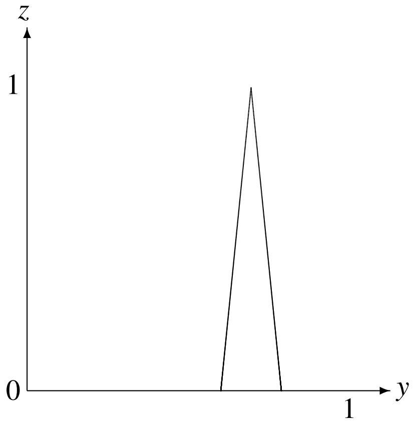


Figure A.8: Triangular secondary membership function of the vertical slice x_2 (figure A.6).

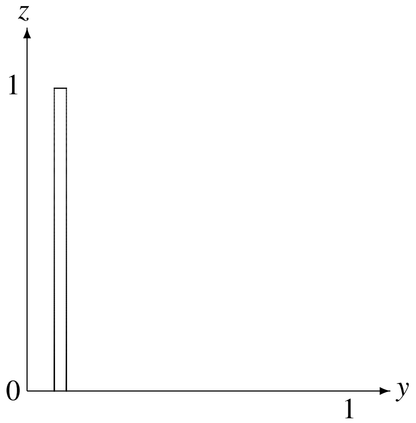


Figure A.9: Rectangular secondary membership function of the vertical slice x_1 (figure A.6). This is an interval type-2 fuzzy set.

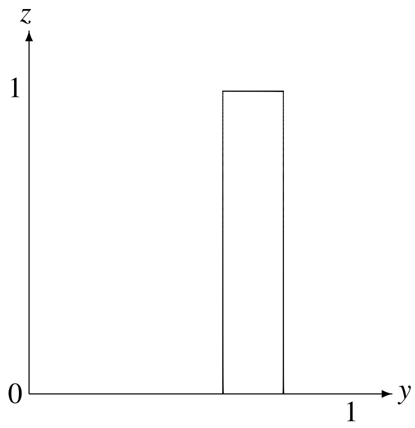
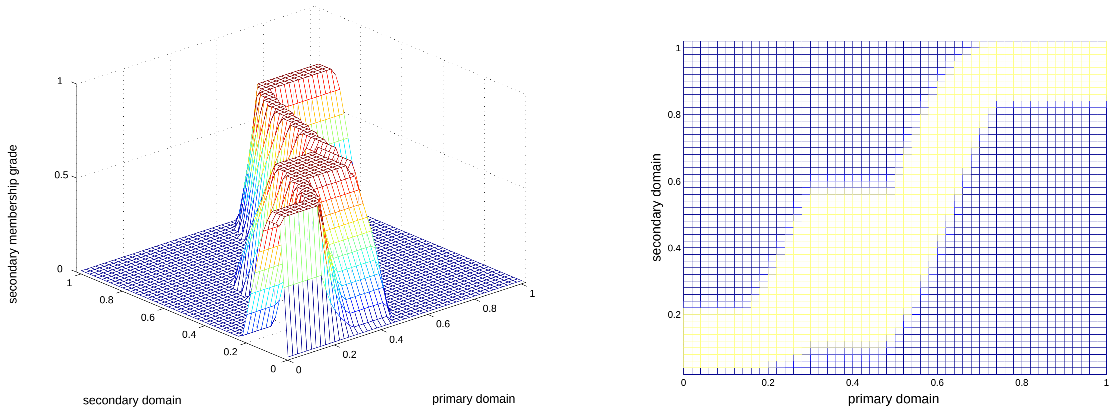


Figure A.10: Rectangular secondary membership function of the vertical slice x_2 (figure A.6). This is an interval type-2 fuzzy set.



(a) 3-D representation

(b) FOU

Figure A.11: Aggregated type-2 fuzzy set created during the inference stage of an FIS.

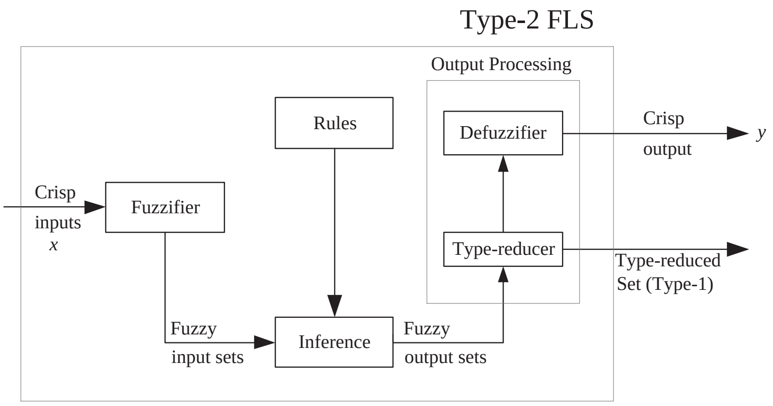


Figure A.12: Type-2 FIS (from Mendel [26]).

Appendix B

String Matching

B.1 Strings

A *string* is simply a list of characters from a pre-defined alphabet. The *string data type* is the way computers represent human writing. Strings might be words, phrases, sentences, and so on. Misspelt words and nonsense words are strings. Since a *space* is simply a character, a phrase such as 'NO PARKING' is regarded as a string of length 10.

Wikipedia define a *string* thus:

> "In mathematical logic, more precisely in the theory of formal languages, and in computer science, a string is a sequence of symbols that are chosen from a set or alphabet.
>
> In computer programming, a string is, essentially, a sequence of characters. A string is generally understood as a data type storing a sequence of data values, usually bytes, in which elements usually stand for characters according to a character encoding, which differentiates it from the more general array data type." [53]

A *substring* is a string which has been 'cut out' from a longer string. For example, if 'NO PARKING' is the string, 'NO' is a substring of length 2, and 'PARKING' is a substring of length 7. 'PARK' and 'KING' are substrings of length 4. 'O PA' and 'ARKI' are also substrings of length 4, nonsense to a human being, but to a computer, perfectly reasonable, well-formed substrings.

B.2 String matching

B.2.1 Exact String Matching

Exact string matching (or *partial string matching*, or just *string matching*) is the technique finding matches for relatively short strings, such as words, within longer strings:

> "String searching algorithms, sometimes called string matching algorithms, are an important class of string algorithms that try to find a place where one or several strings (also called patterns) are found within a larger string or text.
>
> Let Σ be an alphabet (finite set). Formally, both the pattern and searched text are concatenations of elements of Σ. The Σ may be a usual human alphabet (for example, the letters A through Z in English). Other applications may use binary alphabet ($\Sigma = 0,1$) or DNA alphabet ($\Sigma = A,C,G,T$) in bioinformatics." [54]

Gusfield defines the exact string matching problem thus:

> "Given a string P called the *pattern* and a longer string T called the *text*, the **exact matching** problem is to find all occurrences, if any, of pattern P in text T.
>
> For example, if $P = aba$ and $T = bbabaxababay$ then P occurs in T starting at locations 3, 7, and 9. Note that two occurrences of P may overlap, as illustrated by the occurrences of P at locations 7 and 9." [13], page 2

There are several algorithms for exact string matching [54].

B.2.2 Approximate String Matching

Approximate string matching (or *fuzzy*[1] *string matching*) is finding approximate matches for strings, i.e. string matching that accommodates errors.

> "In computing, approximate string matching (often colloquially referred to as fuzzy string searching) is the technique of finding approximate matches to a pattern in a string." [42]

The main application areas of approximate string matching are computational biology, signal processing, and text retrieval [28]. The pattern length depends on the application:

> "The pattern length can be as short as 5 letters (e.g. text retrieval) and as long as a few hundred letters (e.g. computational biology)." [28], page 38

There are numerous algorithms for approximate string matching, all (up to 1999) described by Navarro in his appropriately entitled, "A Guided Tour to Approximate String Matching" [28].

Edit Distance

According to Wikipedia ([42] after [4], page 406),

> The closeness of a match is measured in terms of the number of primitive operations necessary to convert the string into an exact match. This number is called the **edit distance**[2] between the string and the pattern. The usual primitive operations are:[4]
>
> - insertion: cot → co**a**t
> - deletion: co**a**t → cot
> - substitution: co**a**t → co**s**t
>
> These three operations may be generalized as forms of substitution by adding a NULL character (here symbolized by λ) wherever a character has been deleted or inserted:
>
> - insertion: co**λ**t → co**a**t
> - deletion: co**a**t → co**λ**t
> - substitution: co**a**t → co**s**t
>
> Some approximate matchers also treat *transposition*, in which the positions of two letters in the string are swapped, to be a primitive operation. Changing *cost* to *cots* is an example of a transposition. [4]

[1] This is the colloquial sense of 'fuzzy', meaning imprecise, and does not imply use of fuzzy logic.

[2] The edit distance is also known as the Levenshtein Distance, after Vladimir Levenshtein who proposed the measure in 1965 [20].

Error Level

Before attempting to match a pattern with a text, it is necessary to stipulate how many errors (k) are acceptable. In text retrieval it is sensible to keep k low (e.g. 1 or 2). If k is too high, an approximate match can be found for any word. (In the context of computational biology, which deals with strings of DNA that may be thousands of characters long, a larger value of k would be appropriate.)

> "…we have defined the problem of approximate string matching as that of finding the text positions that match a pattern with up to k errors." [28], page 36

For a pattern of length m we can measure the fraction of the pattern that is wrongly matched.

> "…the problem makes sense for $0 < k < m$, since if we can perform m operations we can make the pattern match at any text position by means of m substitutions. …Under these distances, we call $\alpha = \frac{k}{m}$ the *error level*, which given the above conditions, satisfies $0 < \alpha < 1$. This value gives an idea of the "error ratio" allowed in the match (i.e. the fraction of the pattern that can be wrong)." [28], page 38

B.2.3 Contrasting Exact and Approximate String Matching

Exact string matching may be thought of as approximate string matching with an edit distance of 0. Navarro's summary is not concerned with exact string matching:

> "The case $k = 0$ corresponds to exact string matching and is therefore excluded from this work." [28]

Though Navarro does not concern himself with this possibility, approximate string matching may still be used to find exact matches, should they exist. Exact string matching is relatively fast but only finds exact matches. The slower approximate string matching will pick up both approximate and exact matches.

String Matching Using Fuzzy Logic Schneider, Bunke and Kandel [33] have proposed a method employing fuzzy logic to match strings.

> "This paper introduces a new heuristic technique that is based on fuzzy logic to match strings. It is assumed that a document is scanned by some OCR system, and the result was put in a database. The algorithm that is presented here will match the noisy strings in the database against an existing dictionary."

Appendix C

Metadata Schemas

C.1 Metadata

Metadata is information about information. For example, library catalogue data listing author name, publication title, publication date, publisher name, accession number, shelf number and so on, are all items of information that describe the publication concerned. Metadata is recorded systematically using agreed categories (such as author name, publication title, publication date, publisher name, accession number, shelf number and so on). Metadata is useful to close it enables searches and comparisons to be made between different items. (Imagine looking for books on a particular topic in a library without a catalogue.)

One of the most challenging aspects of the digital environment is the identification of resources available on the Web. The existence of searchable descriptive metadata increases the likelihood that digital content will be discovered and used. But metadata has other uses. It is common to distinguish between three basic kinds of metadata:

1. Descriptive metadata helps users find and obtain objects, distinguish one object or group of objects from one another, and discover the subject or contents.
2. Administrative metadata helps collection managers keep track of objects for such purposes as file management, rights management, and preservation.
3. Structural metadata documents relationships within and among objects and enables users to navigate complex objects, such as the pages and chapters of a book [29].

In the context of digital resources, there exists a wide variety of metadata formats reflecting these different uses. Viewed on a continuum of increasing complexity, these range from the basic records used by robot-based Internet search services such as Google, through relatively simple formats like the Dublin Core Metadata Element Set (DCMES) and the more detailed Text Encoding Initiative (TEI) header and MARC formats, to highly specific formats like the FGDC Content Standard for Digital Geospatial Metadata, the Encoded Archival Description (EAD) and the Data Documentation Initiative (DDI) Codebook [37].

Of these perhaps the most well-known metadata initiative is the Dublin Core. The Dublin Core defines fifteen metadata elements for simple resource discovery; title, creator, subject and keywords, description, publisher, contributor, date, resource type, format, resource identifier, source, language, relation, coverage and rights management. One of the specific purposes of DC is to support cross-domain resource discovery; i.e. to serve as an intermediary between the numerous community-specific formats being developed. It has already been used in this way in the service developed by the EU-funded EULER Project and by the UK Arts and Humanities Data Service (AHDS) catalogue. The Dublin Core element set is also used by a number of Internet subject gateway services and in services that broker access to multiple gateways, e.g. the broker service being developed by the EU-funded Renardus Project [27].

One consequence of the wide range of communities having an interest in metadata is that there are a bewildering number of standards and formats in existence or under development. The library world, for example, has developed the MARC formats as a means of encoding metadata defined in cataloguing rules and has also defined descriptive standards in the International Standard Bibliographic Description (ISBD) series. Other domains have defined metadata standards based on implementations of the Standard Generalised Markup Language (SGML) or the Extensible Markup Language (XML). Examples of these are the Encoded Archival Description (EAD), the CIMI Document Type Definition (DTD); an SGML DTD developed by the CIMI consortium [27] and CIDOC-CRM, developed by the International Council of Museums, based on XML/RDF.

C.1.1 The Dublin Core Metadata Element Set

According to Martin Doerr, 'The Dublin Core Element Set can be regarded as the most important metadata standard of the library world and far beyond, to define basic finding aids for electronic resources by a minimal set of semantic fields or "access points".' [8]

Dublin Core is essentially a list of elements (data categories).

Simple Dublin Core The Simple Dublin Core Metadata Element Set [44] consists of 15 metadata elements, namely *Title*, *Creator*, *Subject*, *Description*, *Publisher*, *Contributor*, *Date*, *Type*, *Format*, *Identifier*, *Source*, *Language*, *Relation*, *Coverage*, and *Rights*. Each element is optional and may be repeated. The elements can be used in any order.

Qualified Dublin Core Qualified Dublin Core [44] contains a further 3 elements: *Audience*, *Provenance* and *RightsHolder*.

Dublin Core: Further Extension Over 30 more elements have been added to give "an up-to-date specification of all metadata terms maintained by the Dublin Core Metadata Initiative" [17]. The complete list of elements is: *Abstract*, *Access Rights*, *Accrual Method*, *Accrual Periodicity*, *Accrual Policy*, *Alternative Title*, *Audience*, *Date Available*, *Bibliographic Citation*, *Conforms To*, *Contributor*, *Coverage*, *Date Created*, *Creator*, *Date*, *Date Accepted*, *Date Copyrighted*, *Date Submitted*, *Description*, *Audience Education Level*, *Extent*, *Format*, *Has Format*, *Has Part*, *Has Version*, *Identifier*, *Instructional Method*, *Is Format Of*, *Is Part Of*, *Is Referenced By*, *Is Replaced By*, *Is Required By*, *Date Issued*, *Is Version Of*, *Language*, *License*, *Mediator*, *Medium*, *Date Modified*, *Provenance*, *Publisher*, *References*, *Relation*, *Replaces*, *Requires*, *Rights*, *Rights Holder*, *Source*, *Spatial Coverage*, *Subject*, *Table Of Contents*, *Temporal Coverage*, *Title*, *Type*, and *Date Valid*. [17]

Implementing Dublin Core

Dublin Core may be implemented in (at least) three ways:

1. "Implementations of Dublin Core typically make use of XML and are Resource Description Framework based." [44]
2. Dublin Core may be implemented as a relational database (the current situation with the DMU databases).
3. " ...there's no reason why you couldn't store any data you wanted in a relational model and then ...query that out and put it into an XML document which adheres to the RDF ...document definition." [Simon Coupland]

C.1.2 The CIDOC-CRM

CRM stands for Conceptual Reference Model. The CIDOC CRM is “the culmination of more than a decade of standards development work by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM).” [5] The CRM aims to “solve the problem of semantic interoperability between museum data of various kinds and their relations to archive and library material.” [8] As well as that CIDOC-CRM may be used to inspire good metadata [8].

The basic building block of the CRM is the *class* (or *entity*). CIDOC-CRM is a semantic net, which means that classes may inherit from each other (figure 4.1). Each class inherits from (i.e. is a subclass of) any class to which it has an arrow pointing (superclass); all classes ultimately derive from *E1 CRM Entity*.

As well as classes, CIDOC-CRM defines properties, such as *P4 has time-span*, *P19 was intended use of*, and *P25 moved*, which give *space-time* information and categorise events. Classes can be connected via properties, so creating a rich network which includes the actors and events behind the artifacts in a museum or collection.

Implementing CIDOC CRM

CIDOC CRM is implemented via an XML/RDF database.

C.2 RDF

RDF stands for *Resource Description Framework*, and is code which “provides a standard way of representing metadata.” [31], page 119
This example of RDF code concerns an article about the politician Tony Benn [49].

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="http://en.wikipedia.org/wiki/Tony_Benn">
<dc:title>Tony Benn</dc:title>
<dc:publisher>Wikipedia</dc:publisher>
</rdf:Description>
</rdf:RDF>
```

> “The Resource Description Framework ... enables the creation and exchange of resource metadata as normal Web data. To interpret these metadata within or across user communities, RDF allows the definition of appropriate schema vocabularies (RDFS) ... ” [24]

This example of RDFS records the fact that ‘the resource “horse” is a subclass of the class “animal” ’ [40].

```
<?xml version="1.0"?>

<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xml:base="http://www.animals.fake/animals#">

<rdf:Description rdf:ID="animal">
  <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
```

```
</rdf:Description>

<rdf:Description rdf:ID="horse">
  <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
  <rdfs:subClassOf rdf:resource="#animal"/>
</rdf:Description>

</rdf:RDF>
```

Using RDFS, classes and properties specific to the application can be defined.

> "RDF describes resources with classes, properties, and values. In addition, RDF also need a way to define application-specific classes and properties. Application-specific classes and properties must be defined using extensions to RDF. One such extension is RDF Schema." [40]

> "RDF Schema does not provide actual application-specific classes and properties. Instead RDF Schema provides the framework to describe application-specific classes and properties. Classes in RDF Schema is much like classes in object oriented programming languages. This allows resources to be defined as instances of classes, and subclasses of classes." [40]

In this section of code, taken from the example above, *horse* is defined to be a subclass of *animal*:

```
<rdf:Description rdf:ID="horse">
  <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
  <rdfs:subClassOf rdf:resource="#animal"/>
</rdf:Description>
```

RDF is inextricably linked to XML:

> "RDF was developed at the intersection between the knowledge management world and the library metadata world. It is a graph system layered on top of Extensible Markup Language (XML) . . . " [15], page 2

> "RDF, the Resource Description Framework, was developed by the World Wide Web Consortium (W3C) to create a format for making assertions that leverage the XML format to represent and transport information." [15], page 5

> "The level above XML determines how the information is interpreted, and this is where RDF exists. . . . RDF is a way to express relations between objects, something XML does not allow you to do." [15], page 5

Similarly, RDF syntax derives from XML syntax (so RDF code resembles XML code).

> "RDF is a simple language for expressing data models, which refer to objects ("resources") and their relationships. An RDF-based model can be represented in XML syntax." [50]

RDF goes beyond syntax, and is concerned with semantics. "RDF is a set of rules for creating semantics . . . " [15], page 4

C.2.1 The RDF Database

Triplestores

> “A **triplestore** is a purpose-built database for the storage and retrieval of Resource Description Framework (RDF) metadata. [32]
>
> Much like a relational database, one stores information in a triplestore and retrieves it via a query language. Unlike a relational database, a triplestore is optimized for the storage and retrieval of many short statements called triples, in the form of subject-predicate-object, like “Bob is 35” or “Bob knows Fred”.” [55]

SPARQL RDF data may be queried using SPARQL:

> “SPARQL (pronounced “sparkle”) is an RDF query language; its name is a recursive acronym that stands for **S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage. ...SPARQL allows for a query to consist of triple patterns, conjunctions, disjunctions, and optional patterns. ...Implementations for multiple programming languages exist.” [52]

Serialisation For RDF to be transmitted across the internet, it has to be serialised; it is converted into a linear sequence of bits, to be reconstructed by the destination computer.

> “In computer science, in the context of data storage and transmission, serialization is the process of converting a data structure or object into a sequence of bits so that it can be stored in a file, a memory buffer, or transmitted across a network connection link to be “resurrected” later in the same or another computer environment. When the resulting series of bits is reread according to the serialization format, it can be used to create a semantically identical clone of the original object.” [51]

Communication between Websites Online databases communicate with each other via a *robot*. The robot’s function is to trawl websites or other online data repositories on order to find appropriate data.

C.2.2 RDF Schema

RDFS enables the online sharing of resource metadata as normal web data.

> “The Resource Description Framework ...enables the creation and exchange of resource metadata as normal Web data. To interpret these metadata within or across user communities, RDF allows the definition of appropriate schema vocabularies (RDFS) ... ” [24]

Using RDFS, classes and properties specific to the application can be defined.

> “RDF describes resources with classes, properties, and values. In addition, RDF also need a way to define application-specific classes and properties. Application-specific classes and properties must be defined using extensions to RDF. One such extension is RDF Schema.” [40]

> “RDF Schema does not provide actual application-specific classes and properties. Instead RDF Schema provides the framework to describe application-specific classes and properties. Classes in RDF Schema is much like classes in object oriented programming languages. This allows resources to be defined as instances of classes, and subclasses of classes.” [40]

C.2.3 Fuzzy RDF

When dealing with vague or uncertain information, crisp RDF is inadequate. Fuzzy RDF (mainly type-1) is being developed to overcome the shortcomings of crisp RDF [35] [38] [36]. Li et al. have developed a type-2 version which "can deal with the imprecise knowledge much better" than the type-1 version [21].

C.3 XML

XML (Extensible Markup Language) is a *meta markup language*. A markup language is

> "...a modern system for annotating a text in a way that is syntactically distinguishable from that text. The idea and terminology evolved from the "marking up" of manuscripts, i.e. the revision instructions by editors, traditionally written with a blue pencil on authors' manuscripts. ...Markup is typically omitted from the version of the text which is displayed for end-user consumption." [45]

A more recent example is HTML (Hypertext Markup Language), used to define the layout and appearance of text in Web pages. Again, the markup code is omitted from the displayed version but can be viewed by selecting 'view source' in the browser toolbar.

A *meta* markup language allows us to "create our own markup language ..." [34] XML is a syntax specification, an "agreed-upon protocol for how to create certain kinds of documents." [22] According to Hjelm, XML

> "...is not a markup language; it is a set of rules for creating markup languages. ...XML gives only the rules for how the byte strings should be cobbled together to form a coherent whole, which can be used by a widely spread set of computer programs. ...XML does not say anything about the information itself, only the way it is structured." [15], page 5

At the same time XML is a means of recording data. "Extensible Markup Language (XML) is a data storage toolkit, a configurable vehicle for any kind of information, an evolving and open standard embraced by everyone from bankers to webmasters." [31], page 1

Here is a short yet complete XML document, a record of the 'Foligno' Madonna by Raphael [57]:

```
<?xml version="1.0" encoding="UTF-8" ?>
<painting>
  <img src="madonna.jpg" alt='Foligno Madonna, by Raphael'/>
  <caption>This is Raphael's "Foligno" Madonna, painted in
    <date>1511</date><date>1512</date>.
  </caption>
</painting>
```

The XML markup symbols are known as *tags*, of which '<painting>' and '</painting>' are examples. As is generally the case with tags, they operate as a pair, in this case signifying the beginning and end of a record about a painting.

XML's popularity may be attributed to its 1. allowing easy data exchange, 2. allowing markup languages to be customised, 3. making the data in the document self-describing, 4. allowing for structured and integrated data. [34]

XML Document Preparation XML documents can be prepared in any text editor that saves to plain text format, such as Microsoft's *Notepad* or Unix's *vi*. However a dedicated XML editor may be preferable, offering features convenient to, and in some cases configurable by, the programmer. [31], p 17

XML Schema “An XML Schema describes the structure of an XML document.” [41] XML schemas are beyond the scope of this report.

C.3.1 The XML Database

“An XML database is a data persistence software system that allows data to be stored in XML format. This data can then be queried, exported and serialized into the desired format.” [58]

C.4 Ontologies

“. . . RDF Schema is a way of creating vocabularies.” [15], page 4 These vocabularies are an intermediary between RDF and a true ontology.

> “RDF Schema (variously abbreviated as RDFS, RDF(S), RDF-S, or RDF/S) is an extensible knowledge representation language, providing basic elements for the description of ontologies, otherwise called Resource Description Framework (RDF) vocabularies, intended to structure RDF resources.” [48]

> “In computer science and information science, an ontology is a formal representation of the knowledge by a set of concepts within a domain and the relationships between those concepts. It is used to reason about the properties of that domain, and may be used to describe the domain.” [47]

The difference between a vocabulary and an ontology is that the latter not only defines individual terms but also describes their relationships to the other terms within the same domain. This is done using what are known as ’RDF Triples’ that is to say each term is defined in terms of a subject, predicate and object.

> “The RDF data model [39] is similar to classic conceptual modeling approaches such as Entity-Relationship or Class diagrams, as it is based upon the idea of making statements about resources (in particular Web resources) in the form of subject-predicate-object expressions. These expressions are known as triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object. For example, one way to represent the notion “The sky has the color blue” in RDF is as the triple: a subject denoting “the sky”, a predicate denoting “has the color”, and an object denoting “blue”. . . .
>
> . . . This mechanism for describing resources is a major component in what is proposed by the W3C’s Semantic Web activity: an evolutionary stage of the World Wide Web in which automated software can store, exchange, and use machine-readable information distributed throughout the Web, in turn enabling users to deal with the information with greater efficiency and certainty.” [49]

In their book *Towards the Semantic Web: Ontology Driven Knowledge Management*, Davies, Fensel and van Harmelen ([6], pages 4-5) make the following observations:

- “Ontologies offer a way to cope with heterogenous representation of web resources. The domain model implicit in an ontology can be taken as a unifying structure for giving information a common representation and semantics.”
- “Ontologies are a key enabling technology for the Semantic Web. They interweave human understanding of symbols with their machine processability.”

- "Ontologies were developed in artificial intelligence to facilitate knowledge sharing and re-use."
- "The reason ontologies are becoming popular is largely due to what they promise: *a shared and common understanding of a domain that can be communicated between people and application systems.*"

They outline the scope of ontologies and ontology research,

> "Since the early 1990s, ontologies have become a popular research topic. They have been studied by several artificial intelligence research communities including knowledge engineering, natural-language processing and knowledge representation. More recently, the use of ontologies has also become widespread in fields such as **intelligent information integration**, **cooperative information systems**, **information retrieval**, electronic commerce, and **knowledge management**." [6], page 4

An ontology specifies the terms that can be used in an XML/RDF document. To implement an ontology is to create an XML/RDF document that adheres to the terminology and structure of the ontology.

Appendix D

Fuzzy Photo Project DVD

D.1 Suggested Reading

Table D.1: Selected reading.

REF. NO.	AUTHORS	TOPIC	RELEVANCE
[2]	Cheung et al.	Creating a Fuzzy Inferencing Sys.	Fuzzy MFs & Rules (pp 99–101)
[11]	Greenfield	Uncertainty, Vagueness & Imprecision	Fuzzy Logic in context of AI
[12]	Greenfield & John	Uncertainty of a Type-2 fuzzy Set	Confidence Level
[28]	Navarro	Approximate String Matching	Approximate String Matching
[9]	Doerr et alia.	CIDOC-CRM Introduction	CIDOC-CRM
[5]	Ed. Croft et al.	CIDOC-CRM Definition	CIDOC-CRM

D.2 Knowledge Elicitation Interviews

Table D.2: The knowledge elicitation interviews.

INTERVIEWEE	DATE OF INTERVIEW	DURATION OF INTERVIEW
Stephen Brown	13th May 2010	48 minutes
Kelley Wilder	19th May 2010	60 minutes
Jane Fletcher	20th May 2010	22 minutes
Michael Pritchard	28th May 2010	36 minutes