FuzzyPhoto Interim Report

Author: Stephen Brown
Status: Final
Date: 18/11/13

Contents

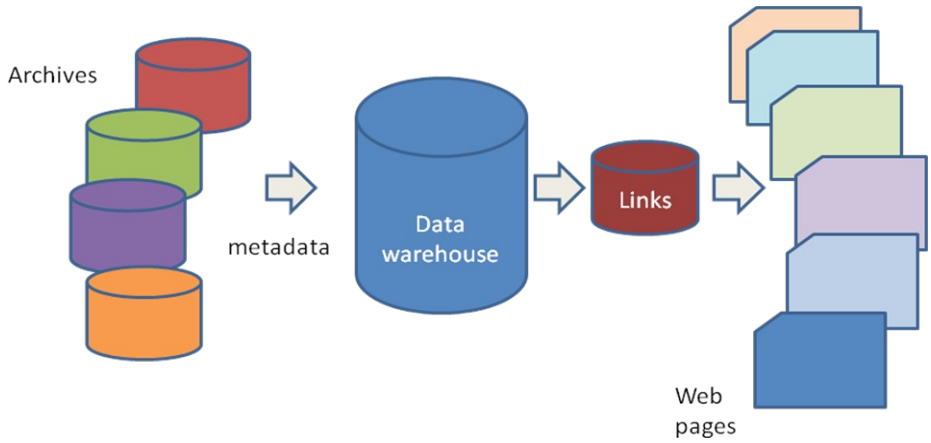
Summary

This is the interim report of the FuzzyPhoto project. FuzzyPhoto is a two year AHRC funded research project (AH/J004367/1) that is developing and testing computer-based “finding aids” that can recommend potential matches between incomplete historical data sets containing imprecise information, even where there is not a precise match.
Half-way through, the project is on schedule. Target records from project partners have been acquired, cleaned, mapped to a common metadata standard and stored in the data warehouse. These along with additional records acquired from other sources bring the total data set to 1.4 million records. A batch loader for semi-automating the process of ingesting records has been developed. Work has started on the mining algorithms that will deliver the record matches aimed for in this project and design of the widget interface to embed these matches in the Web sites of the project partners has also started. Three Research Fellows have been appointed, two of whom have completed their contracts and left the project. Three Advisory Group meetings have been held plus regular weekly project meetings. Four research publications have been generated so far and a project Web site.

Introduction

This is the interim report of the FuzzyPhoto project. FuzzyPhoto is a two year AHRC funded research project (AH/J004367/1) that is developing and testing computer-based “finding aids” that can recommend potential matches between incomplete historical data sets containing imprecise information, even where there is not a precise match, allowing researchers to save considerable time and travel in the early stages of their research when identifying material most likely to be of interest to their studies.

It is centred around a pair of databases that together comprise the single most comprehensive record of British photographic exhibitions in the nineteenth and early twentieth centuries: Photographs Exhibited in Britain (http://peib.dmu.ac.uk) and Exhibitions of the Royal Photographic Society (http://erps.dmu.ac.uk). These early exhibition catalogues were often devoid of pictures, relying instead on written descriptions of the image or artists’ impressions. Image collections are increasingly being published online, creating a timely opportunity to match photographs with their original exhibition catalogue entries without travel to numerous archives. However matching a specific exhibition catalogue reference to a specific image can be a complex and involved process. Photographs were commonly exhibited/published more than once, at different times, with different titles and even by different people, making precise associations between exhibit records and images difficult. This project is developing and testing computer based methods for extracting image object metadata from partners’ collections, storing it in a data warehouse where it can be searched offline for potential matches. The results are stored in a second “links” data base which is interrogatable by a “widget” embedded in partners’ Web sites. When visitors to any one of these sites drill down to the level of an individual image object, the widget will offer them a selection of hyperlinks to potentially related records held in other partners’ collections. Visitors can choose which links, if any, they wish to follow back to the originating partners’ collections. The project will not republish any of the partners’ images or their metadata.



The project outcomes will assist museums, libraries and archives to enhance the value and utility of their collections and of their online services through increased information, improved accuracy and functionality. The project is hosted by the De Montfort University Photographic History Research Centre[1] in collaboration

[1] http://www.dmu.ac.uk/research/research-faculties-and-institutes/art-design-humanities/phrc/photographic-history-research-centre-phrc.aspx

with the Centre for Computational Intelligence[2]. Project partners are Birmingham Central Library, the British Library, the Musee D'Dorsay and the Louvre, Paris, the Metropolitan Museum, New York, the National Media Museum, St Andrews University and the V&A. The project started in November 2012 and ends in November 2014. For further information see the project blog at http://fuzzyphoto.edublogs.org or contact the project leader Professor Stephen Brown sbrown@dmu.ac.uk.

Work plans

Table 1 summarises the project work packages, tasks and their current status.

WP	Title	Lead	Completed
1	**Management**	**Stephen**	
1.1	Staff recruitment and management	Stephen	ongoing
1.2	Progress meetings	Stephen	ongoing
1.3	Finance monitoring and reporting	Stephen	ongoing
1.4	Advisory group recruitment and meetings	Stephen	ongoing
1.5	Annual report writing	Stephen	ongoing
2	**Dissemination**	**Stephen**	
2.1	Press releases	Stephen	ongoing
2.2	Research papers	All	ongoing
2.3	Partner events	Partners	
2.4	Web site (Wordpress blog)	Alex	✓
2.5	Bookmark	Alex	✓
3	**Data acquisition**	**Alex**	
3.1	Review partner collections/select content	Alex	✓
3.2	Metadata schema selection	Alex	✓
3.3	Data ingestion	Alex	✓
3.4	MoU negotiation	Stephen	ongoing
3.5	New contributor negotiation	Stephen	ongoing
3.6	Batch loader	Jethro	✓
4	**Infrastructure**	**Jethro**	
4.1	Hardware/software configuration	Jethro	✓
4.2	Penetration testing	Jethro	
4.3	Sustainability	Stephen	
5	**Data mining**	**Simon**	
5.1	Query expansion	Jethro	✓
5.2	Fuzzy inferencing algorithms	David	ongoing
5.3	Links database	David	ongoing
6	**Interface**	**Simon**	
6.1	Requirements capture	Simon	✓
6.2	Mock-up interface designs	Simon	✓
6.3	Widget implications briefing	Stephen	✓
6.4	Widget implementation agreements with partners	Simon	
6.5	Widget construction	Simon	ongoing
6.6	Interface testing	Stephen	

Table 1. FuzzyPhoto work packages, tasks and progress.

The remainder of this report describes progress against each work package in turn and identifies any issues that have emerged.

WP1 Management

This work package is ongoing.

[2] http://www.cci.dmu.ac.uk/

1.1 Staff recruitment and management

Two Research Fellows were recruited for the start of the project: Dr Alexander von Lunen and Jethro Shell, who subsequently defended his PhD successfully. Dr Lunen was assigned to WP 2 and 3. Dr Shell was assigned to WP 4. Both contracts have now expired. Dr Shell has secured another RF position at De Montfort University and Dr von Lunen is seeking further employment. A third RF was appointed in October 2013, according to plan. The post holder, David Croft, is assigned to WP5. David will defend his PhD in January 2014.

1.2 Progress meetings

Weekly progress meetings have been held since the project began. See Appendix 1 for minutes of these meetings.

1.3 Finance monitoring and reporting

Regular Financial statements have been provided by the Faculty Finance Office. See Appendix 2 for financial statements. The project is within budget.

1.4 Advisory group recruitment and meetings

There have been some changes to the project advisory group members since the original funding proposal was submitted and subsequent to the start of the project, due to changes in the employment circumstances of the individuals concerned. Nevertheless the project successfully recruited and has maintained a full set of advisers and has so far held two advisory group meetings: 6 November 2012 at De Montfort University, Leicester and 5 June 2013 at the V&A. The advisory group is chaired by one of its members, elected by the others. See Appendix 3 for lists of advisory group members and notes on the meetings held to date.

1.5 Annual report

This interim report is the first of two annual reports produced by the project team.

WP 2 Dissemination

This work package is ongoing.

2.1 Press releases

One press statement has been released by De Monfort University.
See Appendix 4 for press release details.

2.2 Research papers

Research publications so far are:
Brown, S., Coupland,S., Croft, D., Shell, J., von Lunen, A. 2013 Where are the pictures? Linking photographic records across collections using fuzzy logic. Paper accepted for Museums and the Web Asia, Hong Kong 9-12 December 2013.
Croft, D., Coupland,S., Shell, J. and Brown, S. 2013 'A Fast and Efficient Semantic Short Text Similarity Metric'. UKCI 2013 (In press)
Croft, D., Coupland, S., Brown, S. 2013 'A hybrid approach to co-reference identification within museum collections'. *2013 IEEE Symposium Series on Computational Intelligence* **(SSCI 2013), Singapore (In press)**
Croft, D., Brown, S., Coupland, S. 2012 'Improving Record Matching Across Disparate Historical Resources.' *Digital Humanities Congress* 6-8 September 2012, The University of Sheffield. Conference abstracts, p.96.

2.3 Partner events

No partner events have yet been held but a promotional bookmark has been distributed at the annual Oracle meeting of museum curators at the Benaki museum in Athens, November 2013. The meeting was attended by 80 delegates.

2.4 Web site

A project Web site was established at http://fuzzyphoto.edublogs.org.. This site acts as a repository for public project documents and a news platform for project developments.

2.5 Bookmark

A promotional bookmark has been designed and printed that provides a brief description of the project and the Web site address. The bookmark was circulated at the Oracle meeting at the Benaki museum, Athens, in November 2013 and will be distributed at the Museums and the Web Asia conference in Hong Kong in December 2013. See Appendix 5 for bookmark visuals.

WP3 Data acquisition

This work package is complete. The elapsed time for this work package was 9 months. The resources required to complete this work package were 32 person-days (partner liaison, etc. not included).

3.1 Review partner collections/select content

The project proposal was supported by a number of institutions, starred in the list below, who agreed to supply catalogue records to the project. These data sets were supplemented by De Montfort's own records (PEIB: Photographs Exhibited in Britain 1839-1865, and ERPS: Exhibitions of the Royal Photographic Society 1870-1915 – ERPS) and data gathered from other collections as shown in table 2.

Partner	Records before cleaning	Records after cleaning
Birmingham City Library	5,513	5,455
British Library	28,974	28,925
CultureGrid	172,148	171,840
ERPS	34,197	34,197
Metropolitan Museum	9,526	9,526
PEIB	20,453	20,453
Musee d'Orsay	46,229	46,228
National Media Museum	8,380	8,380
National Museums Scotland	14,915	14,883
St Andrews University Library	18,620	18,604
Library of Congress	875,267	875,267
Brooklyn Museum	2,352	2,352
National Archives	73,187	71,958
Victoria and Albert Museum	101,538	98,598
		1,406,666

Table 2. FuzzyPhoto source data.

Project partners were visited to review data structures and record management systems and to agree which records to extract (See Appendix 6. Site visit reports). The data from the project partners turned out to be quite heterogeneous in terms of the record management systems used, the way the data were structured, the granularity of data and the degree of separation into different fields. A core set of fields was identified as common to all the data sets, comprising:

- Record ID
- Person name
- Title
- Date
- Dimensions
- Description
- Format

However not all records contained data in every field and how individual field names were interpreted varied depending on the metadata schema employed by the host institution (See Appendix 7. WP 3 report Data Ingestion). Different arrangements were made with each institution to accommodate these variations. WP 3 report Data Ingestion describes the different data structures supplied by each.

3.2 Metadata schema selection

The project originally considered using CIDOC-CRM as an ontology-based data model for the database. However this turned out to be overly complex for the project's needs and circumstances. Since the meta-data delivered by the partners was so diverse, it would have been hard (if not impossible) to comply with CRM, and it would have been necessary to edit the partner's data extensively to comply with CRM. It was decided instead that the Lightweight Information Describing Objects schema (LIDO) offers the best compromise in terms of expressivity, flexibility and implementability. For a more detailed discussion of the rational behind the selection of LIDO for this project see Appendix 7. WP 3 report Data Ingestion.

3.3 Data ingestion

The data from the project partners turned out to be quite heterogeneous and required a good deal of work to unify the various data sets into one schema, necessary to generate the links between the different data. The heterogeneity consists of very different structures of the partners' data, usually introduced by the respective record management software (RMS). Each project partner uses a different RMS with varying export capabilities, limiting the extent to which exporting the data can be customized and therefore made compliant to a specific data model. Therefore, rather than loading the data directly into the chosen metadata (LIDO), temporary tables were created to collate the data and run "clean up" scripts on it. Data was usually delivered as CSV or XML files that were processed manually by importing them into MySQL as specific tables, after which some data cleaning was carried out and then the data was transferred into the chosen metadata schema (LIDO), where some more cleaning was conducted.

It was originally intended to develop a batch loader to achieve this, but variations in the timing of data submission and the very diverse nature of the data made this difficult. The process of importing and cleaning up the data, on the other hand, yielded a very good picture of what would be involved in creating such a tool and a batch loader has subsequently been developed for updating the records. For further details about how the individual data sets were ingested, cleaned and mapped to the LIDO schema, see Appendix 7.

3.4 MoU negotiation

A Memorandum of Understanding was proposed to protect the interests of the project partners. The MoU was intended to prevent commercial exploitation of catalogue records contributed by partners and to prevent retraction of permission to use these records within the project once they had been processed. A specimen MoU is available in Appendix 8. The concept of an MoU was agreed by all the partner institutions. Some were happy to sign the wording of the MoU as suggested in appendix 8. Some requested minor amendments. One partner required a signed contract rather than a MoU, legally binding under French Law. This requirement and to have this agreement set out in both English and French in a form that both institutions would sign introduced significant delays into the process of acquiring this partner's data. A process expected to take two or three months became protracted to a year. At the time of writing this process is virtually complete. The final wording of the contract is being fine tuned in both languages and, in anticipation of imminent signing, the data have been released.

3.5 New contributor negotiation

Since the project aims to find similarities between the records contributed by all the partners, the more records there are, the greater will be the probability that some of them are similar to each other. For this reason the project is open to acquiring additional data sets from institutions outwith the original partnership. Since starting, the following institutions have contributed data:

- The National Archives
- National Museums Scotland

In addition the project has ingested records form a number of institutions that make them available for downloading:

- The US Library of Congress
- Brooklyn Museum
- The Culture Grid

3.6 Batch loader

It was originally intended to develop a batch loader to ingest partners' data, but delays in the data submission by some partners and the very diverse nature of the data made this difficult to achieve and the records have been ingested manually instead. However, building on the experience thus gained, a batch loader has been developed to import additional data from three of the partner organizations. These three were selected because they were able to supply well structured and consistent data amenable to automated processing. The elapsed time for this work was two months. The resources required to complete this work

were 24 person-days. For further details of the batch loader specification and design see Appendix 9. Batch Loader report.

WP 4 Infrastructure

This work package is complete.

4.1 Hardware/software configuration

A MySQL server cluster has been established as the main data repository for the FuzzyPhoto project. Within the cluster a composition of data from ERPS, PEIB and partner organisations is stored. The data comprise meta-data relating to historical photographs within these collections. The cluster also serves to assist in the processing of links between these collections. The structure of the cluster maintains data isolation from external access along with allowing continued, sustained expansion of core structure. Figure 1 shows how the cluster relates to other elements within the process. The main output from the cluster is a links database. Users can access this database via a web server. The links database is periodically updated from the cluster through a defined link.

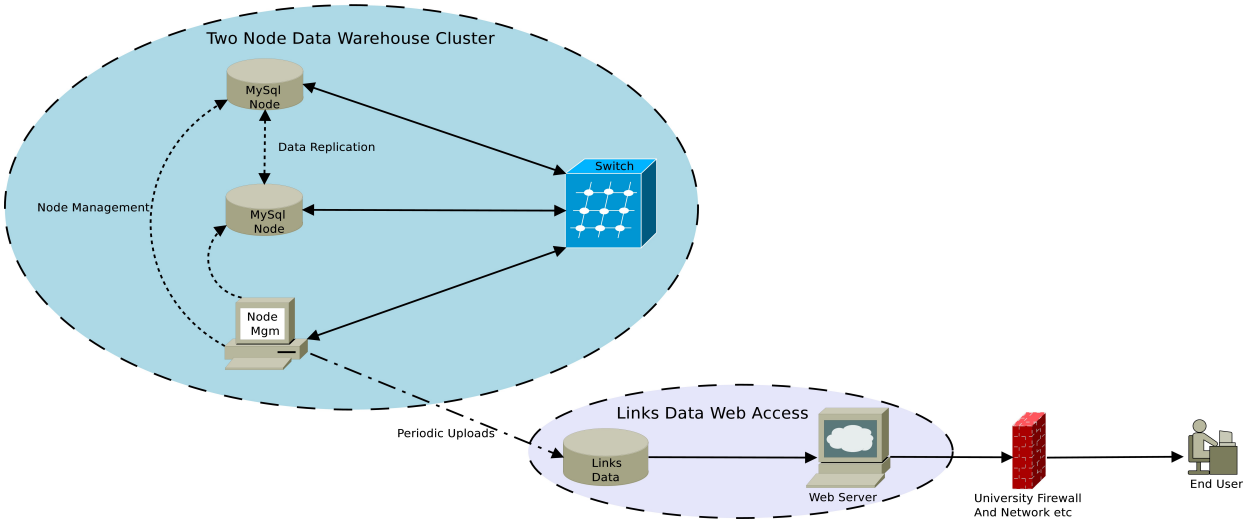


Figure 1. Structure of the FuzzyPhoto server cluster.

For more details of the server cluster see Appendix 10 The FuzzyPhoto MySQLCluster.

4.2 Penetration testing

Since the cluster will be serving content (links) up to partner Web sites it is essential that a high level of security is established and maintained to prevent unauthorised access to partners' data, sites and servers. The physical separation of the Web server from the rest of the cluster network is intended to achieve this. In addition the cluster configuration will be subjected to penetration testing when the links database has been populated.

4.3 Sustainability

The FuzzyPhoto project has funding for just two years but the results are intended to be used by visitors to the partners' sites indefinitely. This requires a robust and well-maintained platform to process and serve up the links to partner Web sites. While the server cluster has been designed to be as robust as possible (see sections 4.1 and 4.2 of this report) server maintenance needs to be addressed in the long term. A number of options are currently under consideration.

WP5 Data mining

5.1 Query expansion

The endpoint of the data ingestion stage is a MySQL database of records in LIDO format. Our aim is to compare these records in order to identify similarities. One of the key fields used to match records is the object title. The individual ERPS record titles each contain very few words, on average 8.1, of which only 5.4 are usable. Such a small amount of text provides little to match against partner records. However we know

that variants of some photographs appear under similar but not identical titles. So by searching for similar words as well as the exact words that appear in the record titles we increase our chances of finding the same photograph even though its title may be different. Query expansion uses semantically similar terms to those in a query to increase the chances of locating matching words (Xu et al., 1996). However before the query terms can be expanded they must first be disambiguated. Many English language words have multiple meanings, for example “fair” can mean “blonde”, “attractive”, “festival” or “equitable”. Therefore when matching an exhibit title such as “Fair Daffodils” the term “fair” has to be disambiguated first. We are using the WordNet Lexical database (http://wordnet.princeton.edu/) to identify synsets of keywords, each of which represents a different meaning for that term, and supplementing this with Part-Of-Speech Tagging (POST). Post applies descriptors to each element in a sentence, such as such as noun, verb, participle, article, pronoun, preposition, adverb, and conjunction, to help disambiguate the words (Voutilainen, 2003). A comparison of alternative software Part-of-Speech taggers indicated that the Stanford POST software (http://nlp.stanford.edu/software/tagger.shtml) is the most accurate when used against a test dataset of ERPS records. For more details about the disambiguation and query expansion techniques we have used see Appendix 11. Word Sense Disambiguation.

5.2 Fuzzy inferencing algorithms

This work package is assigned to the third Research Fellow, appointed from October 2013. As such it has only just started at the time of writing this report. Nevertheless some preliminary work suggests that a fuzzy algorithmic approach can yield interesting results (see Brown, S., Coupland,S., Croft, D., Shell, J., von Lunen, A. 2013 Where are the pictures? Linking photographic records across collections using fuzzy logic. Paper accepted for Museums and the Web Asia, Hong Kong 9-12 December 2013.).

5.3 Links database

Not yet started.

WP6 Interface

6.1 Requirements capture

The “links” data base will be interrogatable by a “widget” embedded in partners Web sites. When visitors to any one of these sites drill down to the level of an individual image object, the widget will offer them a selection of hyperlinks to potentially related records held in other partners’ collections. The FuzzyPhoto widget is a small piece of code that can be inserted into partners’ web pages to display links to corresponding objects in each other's catalogues. Visitors to an object record on a webpage will see a list of hyperlinks suggesting possible matches. Following one of these links will open a new window containing the suggested matching object record within its owner’s web site as shown in figure 2. If there are no suggested matches, no hyperlinks will be shown.

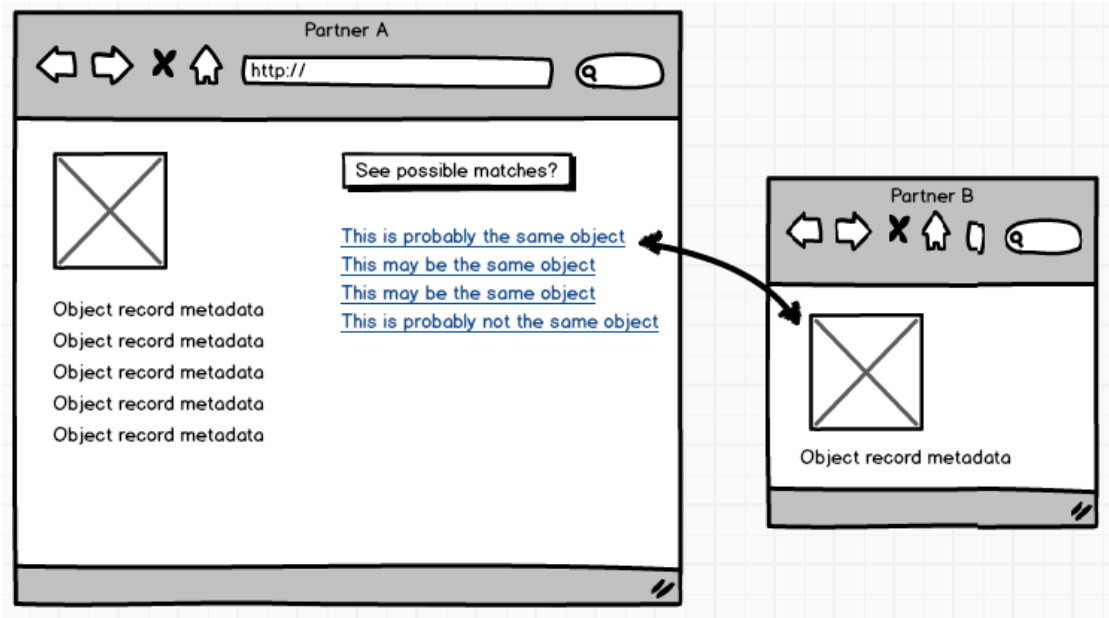


Figure 2. Parent child relationship between object records held by different partners.

6.2 Mock-up interface designs

Partners Web sites and CSS were reviewed to establish how they work and alternative mock up versions were shown to partners at the second advisory group meeting (5 June 2013) to obtain feedback on preferences. The ensuing discussion of these mock-ups generated a number of issues.

At the time of writing, development of working prototypes of widget interfaces is just starting.

6.3 Widget implications for partners

Appendix 12. FuzzyPhoto widget implications for partners, describes in detail how the Widget might work and what important implications the widget may have for partners. The most significant issues are security and sustainability. These issues are addressed in section 4 of this report.

Work remaining

The project is half way through its two-year schedule and on track. Necessary records have been acquired, cleaned, mapped to a common metadata standard and stored in the data warehouse. Work has started on the mining algorithms and widget interface design. This will be developed further and additionally the project will build the links database and penetration test the server cluster to ensure that security is adequate. A major challenge is likely o be gaining acceptance of the IT support functions in the partners' institutions for implementing the widget. For this reason the topic was broached at the second advisory group meeting in June 2013 and it will be the primary focus of the third meeting in February 2014. The FuzzyPhoto widget implications for partners briefing (Appendix 11) was written to help advisory group members to begin the necessary conversations with their management and IT support functions. A further challenge will be long term preservation and sustainability of the links data. A strategy for this will be developed in the second year.

Appendix 1. Project team meeting minutes

FuzzyPhoto Project Meeting 06/07/12

Present: S Brown, S. Coupland

1. Recruitment

Agreed modifications to RA Databases and RA Ontologies.
Reviewed potential recruits: Stephen Matthews, Jethro Shell, Ben Passow, David Croft.
ACTIONS: SB to revise job descriptions and write PAFs by 13/07/12, SC to write Job description/person spec for RA FuzzyLogic by 13/07/12. SB and SC to discuss with David Croft.

2. Project partners

Reported: Delay to project funding means that for the first year of the project Birmingham archives will we inaccessible due to move of Birmingham Library.
Agreed: To find alternative and to use Birmingham collection for verification later in the project.
ACTIONS: SB to find alternative collection, inform AHRC and Birmingham.

3. Press release

Reported: Press release 18/06/12 sent to Jo Griffin, DMU External Relations. Received by SC.
Agreed: Press release to be sent to partners when partner list is finalised and partners invited to issue their own press releases.
ACTIONS: SB to send Press release to partners .

4. Schedule

Reviewed revised schedule. Agreed that start of PDRA could be flexible to some degree to accommodate availability of David Croft.
Agreed research day for SB: Tuesday and for SC: Tuesday am and Thursday am.
Agreed to meet Tuesdays am, weekly.

5. Advisory Board

Agreed draft agenda for first meeting:

- Welcomes
- Review what we are trying to do.
- Review partner expectations.
- Review schedule.
- Q&A
- Knowledge elicitation /user requirement workshops.

ACTION: SB to Doodle poll advisory board members for date in early November.

FuzzyPhoto Project Meeting 28/09/12

Present: S Brown, S. Coupland

1. Management

Informal weekly meetings Tuesdays 9.30-10.00am. First meeting 30 October 2012.
Formal monthly project meetings, first Tuesday of each month 10.00-10.30am.

2. Research days

Stephen: Tuesdays
Simon: Tuesday am and Friday am

3. Web site

ACTION: SC to develop basic site comprising home, partners and staff pages.

4. Equipment

Agreed to consider purchase/loan of laptop for the project.

ACTIONS:

SB to consult David Croft on server spec.

SB to provide desktop PCs/Macs for PDRAs.

5. Publications

ACTIONS: SB to identify spring 2013 GLAM conference for first paper.

6. Schedule

Agreed to bring the start date of the Connections database workpackage forward by 1 month but to retain the original end date.
ACTION: SB to revise schedule and recirculate.

7. Advisory Board

Agreed draft agenda for first meeting.
ACTIONS:
SB to write up and circulate agenda.
SC to book conference room and refreshments
SB to book hotel, dinner and taxis.
SB to draft presentations on project overview and ways of working.
SC to draft presentation on soft computing.

FuzzyPhoto Project Meeting 20/11/12

Present: S Brown, S. Coupland, A von Lunen, J Shell

1. Partner liaison

Limited response from partners.

ACTION: SB to help chase up meeting dates, technical contacts and responses to the MoU.

2. Data acquisition

We need a flexible batch loader able to ingest metadata and parse it into LIDO XML from a variety of different sources without having to be customised each time.

ACTION: JS to draft specification by 27/11/12

3. Data warehousing

ACTION: JS to draft recommendations for hardware required to warehouse the data, by 04/12/12

4. Portland Network

ACTION: JS to make enquiries about the network in Portland to ascertain its capacity by 27/11/12

5. Metadata

ACTION: AvL to investigate tools to convert LIDO to CIDOC CRM, by 27/11/12

6. Project archive

ACTION: JS to set up shared folder on KMD3 for project documents by 27/11/12

7. Research calls

Noted : AHRC and EU calls identified by AvL. Agreed its too early in the project to respond.
ACTION: AvL to ascertain when FP7 ends and FP8 begins.
Noted: SB has initiated conversation with Thierry Gervais at Ryerson:
SB has replied stressing the date range we are working with and enquiring about the tools Thierry is developing.

FuzzyPhoto Project Meeting 27/11/12

Present: S Brown, S. Coupland, A von Lunen, J Shell

1. Partner liaison

Meeting dates fixed for NMeM, Louvre, Musee D'Orsay and National Library, France, V&A, Birmingham and National Museums Scotland. Late January meeting date with St Andrews to be finalised and date set for British Library.

ACTION: AvL to firm up St Andrews date. SB to continue to chase up BL meeting date and confirm time of Birmingham meeting.

2. Data acquisition

Received paper on flexible batch loader drafted by JS.

Agreed: To adopt Individualised approach initially but to proceed to Generalised approach in due course.

ACTION: JS to include roll back function in the spec for the data warehouse.

3. Data warehousing

ACTION: JS to draft recommendations for hardware required to warehouse the data, by 04/12/12

4. Portland Network

Reported: The network in Portland is 1G to the desktop and 10G between buildings. This information will allow load modelling to be conducted.

5. Metadata

Reported: There are various tools to convert LIDO to CIDOC CRM, but none entirely satisfactory.

Agreed: To develop our own if conversion to CIDOC CRM becomes necessary.

6. Research calls

Reported : FP7 ends 2013 and FP8 begins 2014.

7. Fuzzy word sense disambiguation

Reported: There is growing interest in this topic. Published methods are not entirely convincing.

ACTION: SC to continue searching the literature.

8. PHRC 3D Day

Reported: Fuzzy team invited to PHRC 3D Day 4 December, including drinks party at 17.30.

Agreed: To hold brief project meeting at 9.30 as usual.

9. Christmas meeting arrangements

Noted: 11th December meeting may be cancelled if BL meeting is fixed for that day.

18 December meeting cancelled because it coincides with Birmingham Library visit.

No meetings 25 December or 1 January.

First meeting of new year will be 8th January.

Team celebration in Criterion Thursday 20 December 15.00 onwards.

Optional meetings Swan and Rushes Friday 30 November, Friday 7 December , Friday 14 December 17.00.

10. Holiday cards

ACTION: SB to order holiday cards for JS and AvL.

11. Googlel+ Hangout

ACTION: SC to draft user guide.

12. Publicity

ACTION: AvL to draft promotional bookmark.

FuzzyPhoto Project Meeting 04/12/12

Present: S Brown, S. Coupland, A von Lunen, J Shell

1. Partner liaison

Time of Birmingham meeting confirmed.

Received: Report on visit by AvL to NMem 30/1/12. Reported: no problems with proposed fields. No data transfer until MoU is signed but its hoped this can be achieved before Christmas.

Reported: National Museums Scotland concerned that they may not be able to provide images published online and that the quality of their data may not be satisfactory.

ACTIONS: AvL to firm up St Andrews date. SB to continue to chase up BL meeting.

AvL to arrange NMeM data dump in XML preferably before Christmas.

AvL to explain technical requirements to Pam Babes at National Museum Scotland. SB to explain overall requirements to Alison Morrison-Low at National Museum Scotland.

2. Data acquisition

ACTION: AvL to look into how to incorporate URLs from the NMeM Collections Online website into the database.

3. Metadata

ACTION: AvL to continue with LIDO to CIDOC CRM translation investigation.

4. Data warehousing

Received: Report on database model requirements for Data Warehouse by JS.

Agreed: To acquire a new Linux cluster to support the data warehouse and host the links database on the existing Mac server (KMD3)

ACTION: JS to create virtual machine on KMD3 to simulate the Linux cluster pro tem and upload some of DC's data. JS to cost server hardware.

5. Fuzzy word sense disambiguation

Reported: SC is liaising with DC re Wordnet and Fuzzy logic literature.

6. Christmas meeting arrangements

Noted: 11th December meeting may be cancelled if BL meeting is fixed for that day.

18 December meeting cancelled because it coincides with Birmingham Library visit.

No meetings 25 December or 1 January.

First meeting of new year will be 8th January.

Team celebration in Criterion Thursday 20 December 15.00 onwards.

Optional meetings Swan and Rushes Friday 30 November, Friday 7 December, Friday 14 December 17.00.

7. Holiday cards

SB provided holiday cards for JS and AvL.

8. Googlel+ Hangout

ACTION: SC to draft user guide.

9. Publicity

Received: Draft promotional bookmark by AvL.

Agreed: Its excellent.

ACTION: AvL to secure images for the bookmark from partners/Kelley Wilder.

10. Business cards

ACTION: SB to order cards for AvL.

11. Project archive

ACTION: SB to upload key project documents.

FuzzyPhoto Project Meeting 11/12/12

Present: S Brown, S. Coupland, A von Lunen, J Shell

1. Partner liaison

Reported: Visits to BL arranged to meet with John Falconer and Adam Farquhar

ACTIONS: AvL to firm up St Andrews date. AvL to arrange NMeM data dump in XML and chase up MoU preferably before Christmas. SB to explain technical requirements to Pam Babes at National Museum Scotland.

2. Data acquisition

Reported: AvL has written some Javascript to extract data from the NMeM Collections Online website and convert it into a harvestable, structured format.

The Collections Trust Culture Grid service provides free access to online collections data.

Agreed: To search the Culture Grid for PEIB/ERPS photographer names and get an XML dump of any found.

3. Metadata

Reported: AvL is cataloguing LIDO fields and writing XSLT script to extract data from these fields.

4. Data warehousing

Reported: JS has created virtual network on hios own PC with 4 nodes for testing purposes.

ACTION: JS to draft specification and costings for alternative server hardware configurations.

5. Fuzzy word sense disambiguation

Reported: SC is liaising with DC re Wordnet and Fuzzy logic literature.

6. Christmas meeting arrangements

18 December meeting to be held 9.15am prior to Birmingham Library visit.

No meetings 25 December or 1 January.

First meeting of new year will be 8th January.

Team celebration in Criterion Thursday 20 December 15.00 onwards.

7. Googlel+ Hangout

ACTION: SC to draft user guide.

8. Publicity

ACTION: AvL to secure images for the bookmark from partners/Kelley Wilder.

9. Business cards

SB provided cards for AvL.

10. Project archive

Reported: SB upload key project documents to KMD3 project archive.

11. Conference CFP

4th International Conference on the Theory of Information Retrieval (ICTIR 2013), 29 September - 2 October, 2013, Copenhagen, Denmark. http://www.ictir2013.org/cfp.html Deadline for proposals April 2013

ACTION: SC to review previous papers to this conference in February 2013.

FuzzyPhoto Project Meeting 18/12/12

Present: S Brown, A von Lunen, J Shell
Apologies: S. Coupland

1. Partner liaison

Reported: AvL visited BL to review data. BL has 2 cataloguing systems, one on the Web and one internal, both based on a Microsoft SQL server database. No remote access so need to request CSV data dump from their db admin. No images online and no plans to do so. All BL data systems to be converged in future but no known dates for this.

Agreed: To proceed with data in its current form.

ACTIONS: AvL to firm up St Andrews date. AvL to arrange NMeM data dump in XML and chase up MoU preferably before Christmas. SB to chase Pam Babes at National Museum Scotland after Christmas for reply to email explaining technical requirements.

2. Data acquisition

Reported: AvL has extract 4000 recordsfrom the NMeM Collections Online website covering photography related fields.

The Collections Trust Culture Grid service provides free access to online collections data.

ACTION: AvL to search the Culture Grid for PEIB/ERPS photographer names and get an XML dump of any found.

3. Metadata

Ongoing: AvL is cataloguing LIDO fields and writing XSLT script to extract data from these fields.

4. Data warehousing

Received: draft specification and costings for alternative server hardware configurations.

Agreed: provisionally to adopt 3 node solution based on Dell PowerEdge R320 with Mgm 1, A/D – 2A/D – 2 configuration . Estimated cost £3,746.68 plus vat.

ACTION: SC to confirm agreement. JS to prepare F14 purchase request.

5. Fuzzy word sense disambiguation

Ongoing: SC is liaising with DC re Wordnet and Fuzzy logic literature.

6. Christmas meeting arrangements

Team celebration in Criterion Thursday 20 December 15.00 onwards.

No meetings 25 December or 1 January.

First meeting of new year will be 8th January.

7. Googlel+ Hangout

Ongoing: SC to draft user guide.

8. Publicity

Ongoing: AvL to secure images for the bookmark from partners/Kelley Wilder.

9. Conference CFP

4th International Conference on the Theory of Information Retrieval (ICTIR 2013), 29 September - 2 October, 2013, Copenhagen, Denmark. http://www.ictir2013.org/cfp.html Deadline for proposals April 2013

ACTION: SC to review previous papers to this conference in February 2013.

FuzzyPhoto Project Meeting 08/01/13

Present: S Brown, A von Lunen, J Shell, S. Coupland

6. Partner liaison

Reported: V&A MoU signed. Successful visit to Birmingham Library on 18/12/12 (notes attached)

Visits arranged to Musee D'Orsay and Biblioteque Nationale (11 January), V&A (15 January) and British Library (21 January).

ACTIONS: AvL to firm up St Andrews date. AvL to arrange NMeM data dump in XML and chase up MoU. SB to chase Pam Babes at National Museum Scotland after Christmas for reply to email explaining technical requirements. AvL to analyse the publicly accessible Birmingham BPS records to assess what kind of information we can expect. SB to chase Pete James for progress with access to records and MoU.

7. Data acquisition

Reported: Data loss over the Christmas break due to routine re-profiling of AvL's network drive.

ACTION: AvL to request ITMS to reinstate missing folder.

8. Metadata

Ongoing: AvL is cataloguing LIDO fields and writing XSLT script to extract data from these fields.

9. Data warehousing

Received: draft specification and costings for alternative server hardware configurations.

Agreed: Adopt 3 node solution based on Dell PowerEdge R320 with Mgm 1, A/D – 2A/D – 2 configuration plus switch. Cost £3,799.83 inc vat.

ACTION: JS to prepare F14 purchase request for SB approval and submission to Finance.

10. Fuzzy word sense disambiguation

Ongoing: SC is liaising with DC re Wordnet and Fuzzy logic literature.

11. Googlel+ Hangout

Ongoing: SC to draft user guide.

12. Publicity

Ongoing: AvL to secure images for the bookmark from partners/Kelley Wilder.

13. Conference CFP

4th International Conference on the Theory of Information Retrieval (ICTIR 2013), 29 September - 2 October, 2013, Copenhagen, Denmark. http://www.ictir2013.org/cfp.html Deadline for proposals April 2013

ACTION: SC to review previous papers to this conference in February 2013.

14. FuzzyLogic Research Fellow post

Noted: Post is due to start July 2013.

Agreed: Advert to be placed by end of January.

ACTION: SB to request advert to be placed.

FuzzyPhoto Project Meeting 15/01/13

Present: S Brown, A von Lunen, J Shell, S. Coupland

1. Partner liaison

Reported:

Successful visit to Musee D'Orsay and Biblioteque Nationale (11 January) (notes attached).

St Andrews visit scheduled for 18 February . No response from Pam Babes at National Museum Scotland.

ACTION: SB to chase Pete James for progress with access to records and MoU AND Pam Babes for NMS data.

2. Data acquisition

Reported: No progress with recovery of data lost over the Christmas break due to routine re-profiling of AvL's network drive.

ACTION: AvL to chase ITMS request to reinstate missing folder.

3. Metadata

Ongoing: AvL is cataloguing LIDO fields and writing XSLT script to extract data from these fields.

4. Data warehousing

Reported: F14 purchase request for server atc. Submitted to Finance.

5. Fuzzy word sense disambiguation

Ongoing: SC is liaising with DC re Wordnet and Fuzzy logic literature.

6. Googlel+ Hangout

Ongoing: SC to draft user guide.

7. Publicity

Ongoing: AvL to secure images for the bookmark from partners/Kelley Wilder.

8. Conference CFP

4th International Conference on the Theory of Information Retrieval (ICTIR 2013), 29 September - 2 October, 2013, Copenhagen, Denmark. http://www.ictir2013.org/cfp.html Deadline for proposals April 2013

Ongoing: SC to review previous papers to this conference in February 2013.

9. FuzzyLogic Research Fellow post

Ongoing: SB to request advert to be placed.

FuzzyPhoto Project Meeting 22/01/13

Present: S Brown, A von Lunen, J Shell, S. Coupland

1. Partner liaison

Reported:

Successful visit to V&A and British Library (notes attached).

Catalogue data received from from Pam Babes at National Museum Scotland as xls file.

Pete James confirmed data dump will be arranged by Corinna Reyner.

No response yet from Musee D'Orsay.

Awaiting V&A sample data and catalogue standards this week.

No response from BNdF yet.

ACTION: AvL to process NMS data and report on issues. AvL to contact Corrinna Reyner to arrange Birmingham data transfer. AvL to gently prompt Thomas Galifort for progress. AvL to contact Simon Woolf for BL data and chase John Falconer for pictures. SB to send draft MoU to Adam Fraquhar.

2. Data acquisition

Reported: No progress with recovery of data lost over the Christmas break due to routine re-profiling of AvL's network drive.

AvL has downloaded 10Gb of data from the Culture Grid.

Noted: Updating of partner catalogue data needs to be raised as an issue in partner discussions. This could be the focus of future funding bids.

ACTION: SB to escalate ITMS request to reinstate missing folder. AvL to search Culture Grid XML files for usable content. SB/AvL to explore updating of partner catalogue data with partners.

3. Metadata

Ongoing: AvL is cataloguing LIDO fields and writing XSLT script to extract data from these fields.

4. Fuzzy word sense disambiguation

Ongoing: SC is liaising with DC re Wordnet and Fuzzy logic literature.

5. Googlel+ Hangout

Ongoing: SC to draft user guide.

6. Publicity

Ongoing: AvL to secure images for the bookmark from partners/Kelley Wilder.

7. Conference CFP

4th International Conference on the Theory of Information Retrieval (ICTIR 2013), 29 September - 2 October, 2013, Copenhagen, Denmark. http://www.ictir2013.org/cfp.html Deadline for proposals April 2013

Reported: Not directly relevant, decision: not to proceed.

Archives Portal Europe (APEx) conference, Dublin. Deadline for proposals 17 February 2013. Probably no published proceedings, but no conference fee.

ACTION: JS to investigate APEx NoE partners and activities to see if they are relevant to FuzzyPhoto and further possible bids.

8. ICT PSP projects call

Reported: Not sufficiently research oriented and call is too soon, decision: not to proceed.

9. FuzzyLogic Research Fellow post

Ongoing: SB to request advert to be placed.

10. Fuzzy lunches

Agreed: to lunch together on Tuesdays from 29 January onwards.

FuzzyPhoto Project Meeting 29/01/13

Present: S Brown, A von Lunen, J Shell, S. Coupland

1. Partner liaison

Reported: Only NMS data received so far: 14888 records of unstructured data. AvL developing text metric to separate data into separate fields.

ACTION: AvL to telephone partners twice a day for the bext two days to ask for data. After that SB to write to Advisory Group asking for advice on how to accelerate data acquisition.

2. Data acquisition

Reported: No progress with recovery of data lost over the Christmas break due to routine re-profiling of AvL's network drive. Culture Grid data search has been refined to yield circa 2000 items. Data extraction is in progress.

AvL is developing workflow for batch loader based on Culture Grid data analysis. Batchloader will comprise components that can be recombined to create customised profiles for individual partners.

ACTION: SB to chase ITMS request to reinstate missing folder. AvL to search Culture Grid XML files for usable content. AvL to continue data extraction from Culture Grid records and batchloader development.

3. Equipment

Reported: Server delivery date is estimated as mid February.

4. Fuzzy word sense disambiguation

Reported: SC and JS are exploring combination of Bayesian probability and fuzzy logic to disambiguate titles, incorporating lexical descriptions from WordNet.

5. Conference CFP

Archives Portal Europe (APEx) conference, Dublin. Reported: One of three strands is Finding Aids. Agreed: to submit a paper and send delegate to the conference to explore potential synergies and future bid cooperation.

ACTION: AvL to draft abstract by 17 February deadline.

6. Advisory Group meeting

Reported: First Advisory Group meeting was sufficiently under budget to allow next one to be held face-to-face. Science Museum is opening Photography Gallery in July. V&A have offered to host meeting, subject to room availability.

ACTION: SB to canvas availability of members for June date.

7. Next meeting

Monday 4 February 10.00am.

FuzzyPhoto Project Meeting 04/02/13

Present: S Brown, A von Lunen, J Shell, S. Coupland

1. Partner liaison

Reported: Positive phone discussion with Anna Vernon at the BL about wording of the MoU. Resolution anticipated within around a week. No further responses from other partners.

ACTION: SB and AvL to chase.

2. Data acquisition

Reported: ITMS unable to recover data lost over the Christmas break due to routine re-profiling of AvL's network drive. AvL has installed Linux to avoid this problem in future.

AvL is developing workflow for batch loader based on Culture Grid data analysis.

ACTION: AvL to rebuild lost files. AvL to continue data extraction from Culture Grid records and batchloader development.

3. Fuzzy word sense disambiguation

Reported: SC and JS exploration of combination of Bayesian probability and fuzzy logic to disambiguate titles, is ongoing.

4. Conference CFP

Archives Portal Europe (APEx) conference, Dublin. Reported: AvL has produced first draft abstract.

ACTION: AvL to finalise abstract by 17 February deadline.

5. Advisory Group meeting

Reported: Responses to Doodle poll pointing towards 5 June.

ACTION: SB to finalise June date.

6. Next meeting

Tuesday 12 February 9.30am.

FuzzyPhoto Project Meeting 11/02/13

Present: S Brown, A von Lunen, J Shell, S. Coupland

1. Partner liaison

Reported: Draft MoU received from BL. Amendments made and returned to BKL for comment

ACTION: SB to chase.

2. Data acquisition

Reported: Revised data received from V&A. Collections Online data revised. Discussed NMS structure problems.

Action: SC to assist AvL with analysis of NMS data.

3. Server purchase

Delivery estimated in 2 weeks.

4. Fuzzy word sense disambiguation

Reported: SC and JS exploration of combination of Bayesian probability and fuzzy logic to disambiguate titles is ongoing and looking promising.

5. Conference CFP

Received: Draft abstract for Archives Portal Europe (APEx) conference, Dublin by AvL. Some revisions suggested

Action: AvL to revise abstract by 17 February deadline.

6. Advisory Group meeting

Reported: Date confirmed for 5 June. Awaiting estimates from V&A for refreshments.

Action: SB to chase.

7. Digging into Data

Considered Digging into Data Challenge. Agreed to explore possibility of bid.

Action: SB to draft proposal.

8. Next meeting

Tuesday 19 February 9.30am.

FuzzyPhoto Project Meeting 19/02/13

Present: S Brown, A von Lunen, J Shell, S. Coupland

1. Partner liaison

Reported: Second draft MoU received from BL. Confirmation received from Colin Harding that NMeM catalogue data can be released.

ACTION: SB to amend BL MoU and return to BL for comment. SB to chase Colin for NMeM data.

2. Data acquisition

Reported: NMS data analysis ongoing.

3. Fuzzy word sense disambiguation

Reported: JS is writing up code to disambiguate titles. Results of trials on sample data using two words are promising.

4. Conference CFP

Reported: Abstract for Archives Portal Europe (APEx) conference, Dublin submitted by AvL.

Agreed: To draft proposal for UKCI based on fuzzy word sense disambiguation.

ACTION: SC/JS to draft abstract.

5. Advisory Group meeting

Reported: Still awaiting estimates from V&A for refreshments.

Action: SB to chase.

6. Digging into Data

Reported: AvL has identified potential collaborator with text mining background.

Action: SB to draft proposal.

7. Publicity

Reported: BL archival images received for bookmark.

Action: AvL to redraft bookmark. JS to blog about Fuzzy word sense disambiguation technique.

8. Next meeting

Tuesday 26 February 9.30am.

FuzzyPhoto Project Meeting 25/02/13

Present: S Brown, A von Lunen, S. Coupland
Apologies: J Shell

1. Partner liaison

Reported: Signed MoU received from BL. Email received from MuseeD'Orsay indicating that MoU is in preparation and a data sample will be sent within 15 days

2. Data acquisition

Reported: Data is being cleaned and integrated into LIDO framework. Records received so far are:

Culture Grid 172148

Birmingham 5513

Collections online 3357

Metropolitan 9526

NMS 14887

Total 205,431

Reported: BL data team have said response may be slow because the team are too busy to find ways of converting BLOB data type in SQL Server into XML. Agreed: we do not need BLOB data.

Still awaiting revised V&A data and full records.

Still awaiting data from NMeM, 8,500 records promised.

Still awaiting St Andrews data.

ACTION: AvL to tell BL we only need text data as XML, to chase V&A, NMeM and St Andrews if no data by 27/02/13.

3. Fuzzy word sense disambiguation

Reported: Ongoing.

4. Conference CFP

Draft proposal for UKCI based on fuzzy word sense disambiguation pending.

ACTION: SC/JS to draft abstract.

5. Advisory Group meeting

Reported: Estimates from V&A for refreshments received, approximately twice the cost of similar estimates at Leicester. Agreed cost savings on accommodation and evening meals offset increased refreshments costs.

Action: SB to confirm order with V&A. SC, JS and AvL to plan widget presentation on Friday 1 March.

6. Digging into Data

Reported: SB is working on draft proposal.

Action: SB and SC to review draft together. SB to solicit project partners. SB to invite Kelley Wilder to join the project.

7. Publicity

Reported: JS blog about Fuzzy word sense disambiguation technique posted.

Action: AvL to ask Marc Boulay for additional landscape archival images for bookmark.

8. Servers

Reported: Servers delivered. Still awaiting the delivery of ancilliary equipment from Dabs.com.

ACTION: JS to chase delivery of outstanding components. JS/SB to liaise with Frontrunners re swap out of old kit in server rack on Monday 4 March.

Next meeting

Tuesday 5 March 9.30am.

FuzzyPhoto Project Meeting 05/03/13

Present: S Brown, A von Lunen, S. Coupland, J. Shell

1. Partner liaison

Reported: Still no reply from BL data team. Still no data from NMeM.

Still awaiting St Andrews data.

2. Data acquisition

Reported: MuseeD'Orsay data sample of 8 records received. V&A data sent to DMU but unreadable attachment.

Agreed the project can work on a limited data set for now but will need full set by the next Advisory Group meeting on 5 June.

ACTION: AvL to confirm Md'O records are in good order and ask for remainder to be sent asap and to ask V&A to resend data and to check the St Andrews on progress. SC to look at sent email to see if it is possible to access the data.

SB to chase BL

3. Fuzzy word sense disambiguation

Reported: Ongoing. Process chart presented and agreed

ACTION: JS to write up method.

4. Conference CFP

Draft proposal for UKCI based on fuzzy word sense disambiguation pending.

ACTION: SC/JS to draft abstract.

5. Advisory Group meeting 5 June

SC, JS and AvL to plan widget presentation .

6. Digging into Data

Reported: SB circulated draft proposal. Kelley Wilder keen to join the project.

7. Publicity

Received: Draft bookmark design. Agreed more photographs needed.

ACTION: AvL to ask Marc Boulay for additional landscape archival images for bookmark.

8. Servers

Agreed: Physical installation of servers can start this week.

ACTION: JS to liaise with server manager Jeevan Gill

Next meeting

Tuesday 12 March 9.30am.

NB SB will be attending University Research Strategy meeting at 1.00pm so will be unavailable for lunch.

FuzzyPhoto Project Meeting 12/03/13

Present: S Brown, A von Lunen, S. Coupland, J. Shell

1. Partner liaison

Reported: No draft MoU yet from Musee d'Orsay.

2. Data acquisition

Reported: 8,380 records received from NMeM. No records yet from V&A and BL. St. Andrews data delayed, promised for 13/03/13.

Agreed: To use description fields selectively to populate empty title fields and to weight these constructed titles to ensure they don't skew results if they contain misleading terms.

ACTION: SB to chase BL and V&A. AvL to mine description fields for nouns and proper nouns to generate title fields where these are missing.

3. Fuzzy word sense disambiguation

Reported: AvL has identified useful key words (eg "by") that can assist parsing the records.

4. Conference CFP

Reported: JS working on draft proposal for UKCI based on fuzzy word sense disambiguation

5. Advisory Group meeting 5 June

Reported: SC working on javascript based widget presentations within sample of partners' Web pages..

6. Digging into Data

Reported: No response to circulated drafts.

7. Publicity

Received: Additional images promised by Marc Boulay for 13/03/13.

8. Servers

Reported: Servers installed and set up. Awaiting remainder of Dabs order to enable set up of SQL cluster.

ACTION: JS to query delivery with Finance.

Next meeting

Tuesday 19March 9.30am.

FuzzyPhoto Project Meeting 19/03/13

Present: S Brown, A von Lunen

1. Partner liaison

Reported: No draft MoU yet from Musee d'Orsay. SB has telephoned and emailed and left message asking for update.

SB has arranged telephone meeting with Adam Farquhar (BL) for 26/03/13. SB has discussed schedule with Heather Caven (V&A). V&A responded to AvL agreeing to data delivery by end of April.

SB suggested National Library of Wales might be a useful alternative partner.

ACTION: AvL to review National Library of Wales collection to assess whether they could be a useful partner. SB to chase Musee d'Orsay again.

2. Data acquisition

Reported: 18,620 records received from St. Andrews with a number of additional tables providing variants on the main tables. AvL to mine description fields to strip out non- nouns to leave behind terms that could substitute for title fields where these are missing.

3. Fuzzy word sense disambiguation

No report.

4. Conference CFP

No report available on draft proposal for UKCI based on fuzzy word sense disambiguation

5. Advisory Group meeting 5 June

No report on javascript based widget presentations within sample of partners' Web pages.

6. Digging into Data

Reported: No response to circulated drafts.

7. Publicity

Received: draft designs for bookmarks.

ACTION: SB to print colour copies for consideration at next meeting.

8. Servers

No report on remainder of Dabs order to enable set up of SQL cluster.

ACTION: JS to query delivery with Finance.

9. Schedule

ACTION: SB to revise schedule to align it with project developments.

Next meeting

Tuesday 26 March 9.30am.

NB SB will be attending Horizon 2020 workshop at lunchtime.

SC sends apologies for whole meeting.

FuzzyPhoto Project Meeting 25/03/13

Present: S Brown, A von Lunen, Jethro Shell
Apologies: Simon Coupland

1. Partner liaison

Reported: No draft MoU yet from Musee d'Orsay. SB has telephoned and emailed and left message asking for update again.

SB has arranged telephone meeting with Adam Farquhar (BL) for 26/03/13.

Reported: National Library of Wales collection looks relevant.

ACTION: SB to invite National Library of Wales to join the project.

2. Data acquisition

Reported: AvL having to process NMS records by hand because of data format inconsistencies.

3. Fuzzy word sense disambiguation

No developments to report.

4. Conference CFP

Reported: APEX proposal rejected.

No report available on draft proposal for UKCI based on fuzzy word sense disambiguation.

5. Advisory Group meeting 5 June

No report on javascript based widget presentations within sample of partners' Web pages.

6. Digging into Data

Reported: No response to circulated drafts.

Agreed: Project not a realistic prospect.

7. Publicity

Agreed: final version subject to some minor amendments.

ACTION: AvL to circulate to partners including NMS for comment and to request logos again.

8. Servers

Reported: Remainder of Dabs order received. Switch installed, links database set up and secured.

ACTION: JS to finish set up of SQL cluster by 28/03/13.

9. Schedule

ACTION: SB to circulated revised schedule to partners..

Next meeting

Tuesday 9 April 9.30am.

FuzzyPhoto Project Meeting 09/04/13

Present: S Brown, A von Lunen, Jethro Shell, Simon Coupland

1. Partner liaison

Reported: No draft MoU yet from Musee d'Orsay.

Reported: SB invited National Library of Wales to join the project. No response.

ACTION: SB to chase National Library of Wales for reply.

2. Data acquisition

Reported: Received circa 30,000 British Library records apparently containing two data sets.

Agreed: To move data to new server at end of April after V&A data has arrived.

ACTION: AvL to check that the two BL sets are the same.

AvL and JS to devise system for logging data as it is moved across.

Total data acquired to date:

Culture Grid 172148

Birmingham 5513

Collections online 3357

Metropolitan 9526

NMS 14887

NMeM 8380

St Andrews 18620

BL 30000

Total 262431

3. Fuzzy word sense disambiguation

Agreed: JS and CS to prepare report for June Advisory Board.

4. Conference CFP

Reported: Draft proposal for UKCI based on fuzzy word sense disambiguation will be ready by 17 May submission date.

5. Advisory Group meeting 5 June

Agreed outline agenda:

Data acquisition (AvL)

Data conversion and metadata schema (AvL)

Disambiguation and query expansion (JS/CS)

Server configuration and security (JS)

User interface widget prototypes (SC)

Promotional bookmarks (AvL)

ACTION: SB to circulate to partners.

6. Publicity

Reported: New Birmingham logo received requires design changes. Some logos still not available as hi res.

Agreed: Revised layout to accommodate Birmingham.

ACTION: AvL to circulate to partners including NMS for final comment prior to print.

SB to ask Malcolm Daniel for higher level contact at the Met.

7. Servers

Reported: Servers installed and configured. Some minor adjustments required.

Agreed: Penetration testing needs to be carried out when links database is complete.

ACTION: JS to organise pen testing in September 2013.

8. Next meeting

9. Tuesday 16 April 9.30am.

10. Noted: Apologies from SC.

FuzzyPhoto Project Meeting 16/04/13

Present: S Brown, A von Lunen, Jethro Shell
Apologies: Simon Coupland

1. Partner liaison

Reported: No draft MoU yet from Musee d'Orsay.

Reported: No response yet from National Library of Wales.

ACTION: SB to chase National Library of Wales for reply.

2. Data acquisition

Reported: AvL has ascertained which BL data set is the one to use and is converting XML data to a relational table.

JS proposed using Subversion for version control as data is moved across.

3. Fuzzy word sense disambiguation

Reported: JS is developing graph structure of the data and examining ways of automating conversion of foreign words via Bing. Bing service is free for limited amount of data. AvL suggested reducing data volume by using freely available APIs to filter foreign words first.

4. Conference CFP

Reported: JS is working on proposal for UKCI based on fuzzy word sense disambiguation and associated issues such as sparse data corpus.

5. Advisory Group meeting 5 June

Reported: SB has revised the agenda in the light of feedback from Marc Boulay.

ACTION: SB to circulate to partners.

6. Publicity

Reported: Metropolitan and British Library logos received. Design changes discussed to present the logos in alphabetic order and to give DMU and AHRC suitable prominence.

ACTION: AvL to revise design.

7. Servers

Reported: Servers installed and configured.

Next meeting

Tuesday 23 April 9.30am.

NB Dim Sum lunch @ Shanghai Moon

FuzzyPhoto Project Meeting 23/04/13

Present: Stephen Brown, Alex von Lunen, Jethro Shell Simon Coupland

1. Partner liaison

Reported: No draft MoU yet from Musee d'Orsay.

Reported: National Library of Wales declined to join project.

2. Data acquisition

Reported: Converting XML data to a relational table and mapping to LIDO is ongoing.

V&A data expected this week.

3. Fuzzy word sense disambiguation

Reported: Ongoing.

4. Conference CFP

Reported: Proposal for UKCI based on fuzzy word sense disambiguation and associated issues such as sparse data corpus is ongoing. Possibility of second UKCI paper reporting David Croft's results.

ACTION: SC to ask DC for permission to use results.

5. Advisory Group meeting 5 June

Reported: SB has invited Heather Caven t contribute to the agenda.

6. Publicity

Reported: Design changes approved. Still awaiting Musee D'Orsay and Louvre logos

ACTION: SB to chase logos.

7. Interface widget

Tabled: Prototype based on erps site using erps CSS. Issues for discussion at Advisory Group:

If there are no links should the widget be visible?

Should all links load automatically or should there be a link to links?

Should selected links open in a new tab?

What would such a link be called?

Can widget styles be customised (and who will do it?)

8. Connections database

Agreed: migrate data to connections database once it has been converted to LIDO. Circa w/c 13 May.

9. Schedule

Schedule was reviewed in the light of summer holiday plans. No potential problems identified.

10. Next meeting

Tuesday 30 April 9.30am.

FuzzyPhoto Project Meeting 30/04/13

Present: Stephen Brown, Alex von Lunen, Jethro Shell Simon Coupland

1. Partner liaison

Reported: No draft MoU yet from Musee d'Orsay, but an email indicating that it may be imminent. The email also indicated that the Orsay may mistakenly believe that the project requires their images. SB has explained this is not the case.

2. Data acquisition

Reported: V&A data received, circa 101,000 records, including some 34 records relating to post WWII photo albums. Agreed not to include these albums because they are out of scope.

Culture Grid	172148
Birmingham	5513
Collections online	3357
Metropolitan	9526
NMS	14887
NMeM	8380
St Andrews	18620
BL	30000
V&A	101000
Total	363431

3. Fuzzy word sense disambiguation

Reported: Some difficulty combining probabilities. Alternative method is being used, fuzzifying probabilities and using ordered weighted sets.

4. Conference CFP

Reported: David Croft agreed to second UKCI paper reporting his results.

ACTION: SC to write up paper.

5. Advisory Group meeting 5 June

Reported: Draft agenda circulated to members.

6. Publicity

Reported: No response to request to Musee D'Orsay and Louvre for hi res logos

7. Interface widget

Tabled: Updated prototype based on erps site using erps CSS showing expandable list of links.

Next meeting

Tuesday 7 May 9.30am.

FuzzyPhoto Project Meeting 14/05/13

Present: Stephen Brown, Alex von Lunen, Jethro Shell Simon Coupland

1. Partner liaison

Reported: No draft MoU yet from Musee d'Orsay, but Thomas Galifot will be attending the Advisory Group meeting on 5 June.

2. Data acquisition

Reported: Data still being cleaned and mapped to LIDO. This work is on schedule to be finished by 31/05/13. An example of the challenges faced: Birmingham data uses 17 different date formats all of which need to be standardised. Some records contain an additional filed listing exhibitions in which the photographs were exhibited. Agreed: this information should be captured as raw text at least.

3. Fuzzy word sense disambiguation

Reported: Ongoing, looking at ways of tagging parts of speech as well as terminological disambiguation.

4. Conference CFP

Reported: UKCI paper deadline extended to 30 May.

ACTION: DC to present two papers on behalf of the team. DC to obtain costings. SB to investigate School funding.

5. Advisory Group meeting 5 June

Tabled: Draft agenda for discussion.

ACTION: SB to circulate final agenda and attendance list to members. All to prepare presentations using DMU template by 4 June.

6. Publicity

Reported: No response to request to Musee D'Orsay for hi res logos. Louvre have promised logo by the end of the week.

7. Interface widget

Reported: erps widget completed. Agreed: V&A and NMeM to be next examples.

ACTION: SC to continue widget development.

Next meeting

Tuesday 21 May 9.30am.

FuzzyPhoto Project Meeting 18/05/13

Present: S Brown, Simon Coupland, A von Lunen, Jethro Shell

1. Partner liaison

Reported: Email from Musee d'Orsay indicating they now grasp the project essentials.

Reported: No logo yet from Lovre

ACTION: SB to chase Louvre for logo.

2. Data acquisition

Reported: Data acquisition and cleaning mostly complete (apart from Musee D'Orsay)

ACTION: AvL to write up work package report. JS to upload data to data warehouse.

3. Fuzzy word sense disambiguation

Reported: Ongoing. How the part of speech tagger will work will depend on the yet to be developed search algorithm.

4. Conference CFP

Reported: Proposal for UKCI based on fuzzy word sense disambiguation and associated issues such as sparse data corpus abandoned. DC is working on submission based on his preliminary results.

Agreed: SB to draft proposal outline for Museums and the Web Asia (MW Asia, Hong Kong, Dec 2013). Deadline for 500 word abstract is 30 June, full paper by 30 September.

SC to draft proposal for World Congress on Computational Intelligence (WCCI, Beijing, July 2014). Deadline for abstract is Dec 20 2013, full paper by April 15 2014.

SB to scope DRHA and alternatives.

ACTION: AvL to recirculate unsuccessful Dublin 'Building Infrastructures for archives in a digital world' abstract.

5. Publicity

ACTION: AvL to revise design in light of feedback received from Advisory Board members and to reinstate Musee D'Orsay Logo. NB Still awaiting Louvre Logo

6. Advisory Group meeting 5 June actions

SB circulate Advisory meeting slides and notes (done)

Susanna Brown to provide National Portrait gallery and Marthe Weiss contacts

SB to provide breakdown of risks to partners re widget installation for each type of widget and supporting arguments.

SB to draft widget testing protocols.

7. Next meeting

Tuesday 25 June 9.30am. (Apologies SC)

8. Summer meeting schedule

23 July 9.30

27 August 9.30

FuzzyPhoto Project Meeting 23/07/13

Present: S Brown, Simon Coupland, A von Lunen, Jethro Shell

1. Partner liaison

Reported: Email from Musee d'Orsay requesting widget implementation details

Reported: Louvre logo received.

ACTION: SB to respond to D'Orsay widget questions.

2. Data acquisition

Reported: Data acquisition and cleaning mostly complete (apart from Musee D'Orsay)

ACTION: AvL to write up first draft of work package report by 31 July 2013. JS to upload data to data warehouse.

3. Fuzzy word sense disambiguation

Agreed: To use standard third party part of speech taggers, eg. Stanford.

ACTION: JS to write up first draft of work package report by 8 August.

4. Conference CFP

Reported: DC proposal for UKCI based on his preliminary results accepted.

Agreed: Project to fund David's attendance at UKCI, Guildford, up to £500.

Reported: SB submitted proposal outline for Museums and the Web Asia (MW Asia, Hong Kong, Dec 2013). Deadline for full paper by 30 September.

ACTION: SC to draft proposal for World Congress on Computational Intelligence (WCCI, Beijing, July 2014). Deadline for abstract is Dec 20 2013, full paper by April 15 2014.

ACTION: SB to scope DRHA and alternatives.

5. Publicity

ACTION: AvL to revise design in light of feedback received from Advisory Board members and to reinstate Musee D'Orsay Logo.

6. Advisory Group meeting 5 June follow up

Reported: Susanna Brown provided National Portrait Gallery contact: Clare Freestone. SB is discussing with Clare the possibility of including the NPG photographic image record metadata in FuzzyPhoto.

Reported: Marthe Weiss (V&A) agreed to promote FuzzyPhoto at Oracle (November).

ACTIONS:

SB to continue discussion with Clare Freestone, NPG.

SB to provide Marthe Weiss with project information in October 2013.

SB to provide breakdown of risks to partners re widget installation for each type of widget and supporting arguments.

SB to draft widget testing protocols.

7. Batch Loader

Agreed to build at least one batch loader, based on CSV data input, by early October.

ACTIONS:

JS to spec and build data ingester, to be operational at fixed times.

AvL to define automated data cleaning requirements and to test David's code to find out where/why it fails.

8. AHRC Big Data cfp

Reported: SB attended AHRC workshop on 25 June re AHRC cfp for Big Data projects, deadline 12 September. The AHRC has a total of £4m capital funding available under this call. This forms part of the Digital Transformations in the Arts and Humanities theme, and proposals must both compliment and add value to the core aims of the theme, and previous research funded through the theme. The call aims to address some of the challenges that arise from working with big data, as well as asking interesting questions of data, and producing innovative and creative assets for future Arts and Humanities research. As a main output, these research projects should produce a tangible asset that sustains beyond the life of the project. Funding for either smaller projects of up to £100k, or larger projects of up to £600k is available on a fEC basis, with the AHRC meeting around 80% of the fEC. Awards should last for a maximum of 15 months and will be expected to start on 1 January 2014 and finish by 31 March 2015.
http://www.ahrc.ac.uk/Funding-Opportunities/Pages/Big-Data-Projects-Call.aspx

ACTION: SB to follow up contacts made at the workshop:

Dr Valerie Johnson, Head of Research, The National Archives

Colin Johnson (C.G.Johnson@kent.ac.uk), Sub-Dean (Graduate Studies), Faculty of Sciences Senior Lecturer, School of Computing University of Kent, Canterbury, UK
http://www.cs.kent.ac.uk/people/staff/cgj

9. Next meeting

27 August 9.30

FuzzyPhoto Project Meeting 27/08/13

Present: S Brown, Simon Coupland, A von Lunen, Jethro Shell

1. Partner liaison

Reported: Despite reminders to the Musee d'Orsay they have not yet sent a revised MoU.

2. Data acquisition

Reported: National Archives have supplied a CSV dump of their photographic records.

ACTION: AvL to process NA records. Also to consult with David Croft re including sources used by David for his thesis, eg. Library of Congress.

Reported: Data Ingestion report received.

ACTION: AvL to make minor amendments and to incorporate earlier paper on metadata selection.

3. Fuzzy word sense disambiguation

Reported: Draft report received and corrections agreed.

ACTION: JS to upload final version to the Blog and to the project server archive.

4. Conference CFP

Reported: Proposal UKCI paper accepted as a poster.

ACTION: DC to draft poster.

Reported: UKMW Asia paper accepted.

ACTION: SB to draft paper.

Reported: Other conference possibilities are Digital Humanities (Lausanne) and WCCI.

ACTION: AvL to draft proposal based on Data ingestion report.

5. Publicity

ACTION: SB and AvL to finalise the bookmark design.

Agreed: Final versions of project reports and minutes to be placed on project Blog.

ACTION: All.

6. Advisory Group meeting 5 June actions

Reported: SB circulated breakdown of risks to partners re widget installation.

ACTION: SB to draft widget testing protocols.

7. Project work packages

Agreed: List of work packages and tasks (attached).

ACTION: Future reports to reference this agreed list.

8. Batch Loader

Agreed: Example batch loaders to be built for V&A, BL and Birmingham

ACTION: JS.

9. Next meeting

Tuesday 3 September, 9.30.00am

FuzzyPhoto Project Meeting 17/09/13

Present: S Brown, Simon Coupland, A von Lunen, Jethro Shell

1. Partner liaison

Reported: Susanna Brown, Curator of Photographs at the V&A to replace Heather Caven on the Advisory Board. Heather is moving to National Museums Scotland.

2. Data acquisition

Reported: 875,267 Library of Congress records and 2352 Brooklyn museum records ingested.

Reported: Data Ingestion report updates ongoing.

ACTION: AvL to process National Archive, PEIB and ERPS records.

3. Fuzzy word sense disambiguation

Reported: Final version of the report uploaded to the Blog and to the project server archive.

4. Conference CFP

MWAsia paper submitted for peer review.

5. Publicity

Reported: Bookmark design finalised.

ACTION: AvL to secure quotes from John E Wright and DMU Print Services for 200 print run and circulate design to partners for final comment.

6. Widget development

Reported: LAMP server set up for widget development. .

7. Batch Loader

Reported: Prototype batch loaders under development. Can detect and autoload clean CSV and XML data.

ACTION: JS.

8. Next meeting

Tuesday 24 September, 9.30.00am

FuzzyPhoto Project Meeting 24/09/13

Present: S Brown, A von Lunen, Jethro Shell

Apologies: Simon Coupland,

1. Partner liaison

Reported: Malcolm Daniel is moving to the Museum of Fine Art, Houston in December. SB has asked if the MFA collection can be added to FuzzyPhoto.

2. Data acquisition

Reported: National Archives records being processed.

Agreed: Incoming Frontrunner Manish Patel to be trained on ingestion procedures using PEIB and ERPS records.

3. Conference CFP

Reported: JS and CS developing idea for WCCI based on DC's date and person metrics

4. Publicity

Reported: Best bookmark print quote received from DMU Print Services. No unfavourable comments receivedfrom partners on the final design.

ACTION: SB to authorise F12. AvL to arrange printing.

5. Batch Loader

Reported: Prototype developed for ingesting V&A and Birmingham data into tables

ACTION: JS to continue development to develop functionality for cleaning data and converting it to LIDO.

6. Further bids

Reported: HEIF FuzzyFinder proposal successful. Awarded £22,000. Project start date January 2014 although may be brought forward.

ACTION: SB to arrange project start.

7. Next meeting

Tuesday 1 October, 9.30.00am

FuzzyPhoto Project Meeting 01/10/13

Present: S Brown, Simon Coupland, A von Lunen, Jethro Shell, David Croft

1. New staff welcome

David Croft welcomed to the project team.

2. Data acquisition

Reported: 71,958 National Archives records added.

3. Conference CFP

Reported: JS and CS developing idea for WCCI based on DC's date and person metrics, deadline 21 December. No news on MW Asia paper.

4. Publicity

Reported: Bookmark ready for test print.

ACTION: AvL to arrange printing.

5. Batch Loader

Reported: Prototype developed for ingesting V&A and Birmingham data into tables now includes St Andrews.

ACTION: JS to continue development to develop functionality for cleaning data and converting it to LIDO. DC to prioritise this over Links Database development.

6. Further bids

Reported: Budget awarded to HEIF FuzzyFinder project is less than amount bid for.

ACTION: SB to investigate feasibility of running the project on the reduced budget.

7. Widget development

Agreed: discoverd links will be directional, ie. 1-way. DMU will hold the links database and partners will be given insititutional-specific views that allow them to see only links to their own seed records.

ACTION: SC to continue development.

8. Penetration testing

Reported: Server structure test successful.

ACTION: JS to discuss Web application testing with Clinton Ingrams and ITMS.

9. Links database

Reported: Rough mock-up started comprising 2 tables.

ACTION: DC and SC to complete in 2 weeks.

10. Next meeting

Tuesday 8 October, 10.00am

NB new start time to accommodate Simon's teaching timetable.

FuzzyPhoto Project Meeting 08/10/13

Present: S Brown, Simon Coupland, A von Lunen, Jethro Shell, David Croft

1. Data acquisition

Reported: Manish Patel is processing PEIB and ERPs records. AvL is cleaning up NMS data.

ACTION: AvL to write up WP report.

2. Conference CFP

Reported: No news on MW Asia paper. WCCI paper in progress.

3. Publicity

Reported: Bookmark sent for printing.

4. Batch Loader

Reported: Prototype developed for ingesting V&A and Birmingham data into tables now converts to LIDO via an intermediate table. Process data is messy.

ACTION: JS to continue development. DC to get advice from Kelly Wilder about how to fit unusual processes (eg. Photoglyphs) into the schema.

5. Further bids

Reported: No resolution of HEIF FuzzyFinder project budget shortfall yet.

ACTION: SB to continue to investigate feasibility of running the project on the reduced budget.

6. Widget development

Ongoing.

ACTION: SC to continue development.

7. Penetration testing

Reported: Potential PHP vulnerability.

ACTION: SC to investigate whether PHP is necessary

8. Links database

Reported: ongoing

ACTION: DC and SC to complete in 1 week.

9. Next meeting

Tuesday 15 October, 10.00am

FuzzyPhoto Project Meeting 15/10/13

Present: S Brown, Simon Coupland, A von Lunen, Jethro Shell, David Croft

1. Partner liaison

Reported: Draft contract received from Musee d'Orsay. Some corrections made and contract returned to Paris and also copied to Christian Taylor (European contract lawyer) for full translation.

2. Data acquisition

Reported: Manish Patel is processing PEIB and ERPs records. AvL is cleaning up NMS data.

ACTION: AvL to circulate WP report.

3. Conference CFP

Reported: First review of MW Asia paper received: Excellent review, no corrections. WCCI paper ongoing.

4. Publicity

Reported: Bookmark printed.

ACTION: SB to remind Marte Weiss about circulation of the bookmark at the November Oracle meeting. AvL to save the bookmark master file to the project server on KMD3.

5. Batch Loader

Reported: Batchloader is substantially complete and WP report is being written up.

ACTION: JS to continue writing up WP report and advise partners that we expect them to provide complete data sets whenever they update records, not just the updated/new records. DC to get advice from Kelly Wilder about how to fit unusual processes (eg. Photoglyphs) into the schema.

6. Further bids

Reported: HEIF FuzzyFinder project budget revised to include PI contribution.

ACTION: SB to meet with Tim Watson to agree project schedule.

Agreed: Not to respond to AHRC CfP for Follow-on Funding for Impact and Engagement until Easter 2014.

7. Widget development

Ongoing.

ACTION: SC to continue development.

8. Penetration testing

Reported: Trial version flat file of up to 1 million records tested without noticeable lag.

ACTION: SC to continue testing up to 2 million records.

9. Links database

Reported: Trial version built. Extra RAM needed.

ACTION: SC to investigate security implications. DC/JS to draft F14 for additional RAM.

10. Next meeting

Tuesday 22 October, 10.00am

FuzzyPhoto Project Meeting 22/10/13

Present: S Brown, Simon Coupland, A von Lunen, Jethro Shell, David Croft

1. Partner liaison

Reported: No progress on Musee d'Orsay contract.

2. Data acquisition

Reported: Manish Patel is processing PEIB and ERPs records. AvL is cleaning up NMS data.

ACTION: AvL to circulate WP report and show Manish how to import XML data.

3. Conference CFP

Reported: No further reviews of MW Asia paper. WCCI paper ongoing. December deadline likely to be extended to January.

4. Publicity

Reported: Marte Weiss not attending Oracle meeting. Bookmark master file saved to the project server on KMD3.

ACTION: SB to contact Oracle organisers about project publicity.

5. Batch Loader

Reported: V&A confirmed agreement to provide complete data sets whenever they update records, not just the updated/new records.

ACTION: JS to continue writing up WP report and circulate it. JS to configure Web server to accept incoming files from partners via SSH and FTP on the virtual server to pull files across.

6. Further bids

Reported: HEIF FuzzyFinder project research post approved

ACTION: SB to arrange advert and interviews.

7. Widget development

Trial version flat file of up to 2 million records tested without noticeable lag.

Agreed: To proceed with flat file rather than RDB and PHP.

ACTION: SC to continue development.

8. Penetration testing

ACTION: SC to test iFrame loading.

9. Links database

Reported: Extra RAM ordered.

10. Next meeting

Tuesday 29 October, 10.00am

FuzzyPhoto Project Meeting 05/11/13

Present: S Brown, S Coupland, A von Lunen, David Croft

1. Partner liaison

Reported: Draft Musee d'Orsay contract received from Christian Taylor. SB has amended it and returned it for translation.

2. Data acquisition

Reported: Orsay data includes a number of apparently duplicate records. Agreed: To not process duplicate records into LIDO. Data copied to KMD3 shared directory and to FuzzyPhoto account on AvL's own PC.

ACTION: AvL to circulate WP report , finish importing Orsay data and NMS data.

3. Publicity

Reported: 80 project project bookmarks sent to the Oracle meeting at the Benaki museum in Athens.

4. Further bids

Reported: No news yet on Big Data proposal.

5. Links database

Reported: outputting results using process data only.

ACTION: DC to continue algorithm development.

6. Server cluster

Reported: Extra RAM not yet arrived.

Server documentation received.

7. Widget

Agreed: Draft widget application to be running by Christmas 2013 on specimen data.

ACTION: SC

8. Publications

Possible journal publications: Foundations and Practice of Information Retrieval (DC), Literary and Linguistic Computing (SB), Information Sciences (SC)

9. Departing staff

Reported: Alex and Jethro have now left the project due to expiry of their contracts. Agreed they have both done a tremendous job.

10. Next meeting

Tuesday 12 November, 10.00am

Appendix 2. Project financial statement

Project Finances

Code: 10.1.089

Project title: Fuzzy Photo

Funding Source: AHRC
Duration: 24 Months (01/11/2012 - 30/10/2014)
Manager: Stephen Brown

	100% **Proforma Budget**	80% **Award Budget**	80% income -100% exp	**Revised Budget**	9 months 1-12 2012 **Budget Year 1**	Virement **Year 1**	Virement II **Year 1**	12 months 1-12 2012 **Actual Year 1**	12 mths 1-12 2013 **Budget Year 2**	3 mths 1-12 2014 **Budget Year 3**
Income Details	Date:	Date:		Date:						
1514 - RC/UK Charity Misc ext Income	-£313,757.00	-£251,005.60	-£251,005.60		-£117,168.22			-£150,947.43	-£99,415.65	-£17,348.11
1517 - Invoice Accruals RGeC								£16,705.59		
Staff Costs										
2141 - Ac Pay Research Gross	£112,878.00	£90,485.00	£112,878.00		£42,329.25			£46,300.18	£45,181.70	£11,295.43
2142 - Ac Pay Research Super								£6,528.34		
2143 - Ac Pay Research NI								£3,572.35		
2190 - Ac Pay Recharges (Stephen)	£68,937.00	£55,485.00	£68,937.00	£39,943.00	£14,978.63			£26,910.16	£33,621.47	£8,405.37
2191 - Ac Pay Recharges					£0.00	£0.00	£0.00	£0.00		
2190 - Ac Pay Recharges (Simon)				£28,994.00	£10,872.75			£0.00		
Non-Staff Costs										
3230 - Computer Equipment	£5,762.00	£4,638.00	£5,762.00		£0.00	£3,628.80		£3,614.36	£2,147.64	£0.00

3413 - Staff Travel - Other	£20,615.00	£16,592.00	£20,615.00		£9,563.75	-£140.00	-£130.00	£1,889.46	£18,464.84	£0.00
3601 - Hospitality						£140.00	£130.00	£260.70		
Contribution	-£105,565.00	-£83,805.60	-£42,813.60	£68,937.00	-£39,423.85			-£45,166.29	£0.00	£2,352.69
	-50.71%	-50.12%	-20.56%	172.59%	-50.71%			-50.71%	0.00%	11.94%

Appendix 3. Advisory group

Advisory group original composition as per project proposal:

Heather Caven, Head of Collections Management, Victoria and Albert Museum
Marc Boulay, Photographic Archivist, University of St Andrews Library *elected chair*
John Falconer, Lead Curator of Visual Arts, British Library
Thomas Galifot, Photography Curator, Musee d'Orsay, Paris
Colin Harding, Curator of Photographic Technology, National Media Museum
Pete James, Head of Photographs, Birmingham Central Library
Professor Robert John, Director of the Centre for Computational Intelligence, DMU
Dominique de Font Reaulx, Conservateur, Musee du Louvre, Paris
Malcolm Daniel, Curator in Charge, Photography, Metropolitan Museum, New York
Professor Roger Taylor, Professor Emeritus, Photohistory, DMU
Stephen Brown *ex offcio*
Simon Coupland *ex officio*

Professor Robert John left DMU before the start of the project and resigned from the advisory group.
Heather Caven left the V&A in September 2013 and was replaced by Susanna Brown.
Malcolm Daniel moved from the Metropolitan Museum to the Museum of Fine Arts, Houston in December 2013.

Advisory group composition at December 2013

Susanna Brown, Curator, Photographs, Victoria and Albert Museum
Marc Boulay, Photographic Archivist, University of St Andrews Library, *elected chair*
John Falconer, Lead Curator of Visual Arts, British Library
Thomas Galifot, Photography Curator, Musee d'Orsay, Paris
Colin Harding, Curator of Photographic Technology, National Media Museum
Pete James, Head of Photographs, Birmingham Central Library
Dominique de Font Reaulx, Conservateur, Musee du Louvre
Malcolm Daniel, Curator in Charge, Photography, Museum of Fine Arts, Houston, Texas
Professor Roger Taylor, Professor Emeritus, Photohistory, DMU
Stephen Brown *ex offcio*
Simon Coupland *ex officio*

FuzzyPhoto

Project Advisory Committee

Terms of reference

1. The role of the Project Advisory Committee(PAC) is to give advice on the strategic development of the project, especially in relation to:

- The validity and coherence of the project, including the research methods and project outcomes;
- Ways of making the results of FuzzyPhoto more useful for, and relevant to, users in the GLAMs (Galleries, Libraries, Archives and Museums) and research fields;

- Parallel developments of relevance to FuzzyPhoto;
- Potential areas of collaboration with other projects;
- Ethical and commercial issues or implications, arising from the project;
- Dissemination plans and opportunities.

2. Recommendations of the Advisory Group will be considered by the Project Management Group and reported in the Annual Reports to the AHRC, but are not binding on the project.

Membership

1. The Advisory Committee will comprise members representing the project partners including the project PI and CI.
2. Additional members may be co-opted as appropriate.
3. The PAC will elect a chair.
4. The Project PI will act as secretary to the Committee.

FuzzyPhoto Advisory Committee 6 November 2012

De Montfort University, Leicester
Gateway House 4.38

Minutes

Present: Marc Boulay, Professor Stephen Brown, Heather Caven,
Dr Simon Coupland, John Falconer, Dominique de-Font-Reaulx, Colin Harding, Professor Roger Taylor
Apologies: Malcolm Daniel, Thomas Galifot, Pete James

In attendance: David Croft, Dr Alexander von Lunen, Jethro Shell.

Welcomes and Introductions

Dr Gerard Moran PVC/Dean of Faculty of Art, Design and Humanities, welcomed committee members to De Montfort University and thanked them for their contribution to the project.
Individuals introduced themselves briefly. Details of job titles and institutions are included in the Fuzzy Photo Contacts sheet circulated separately.
Agreed: Marc Boulay to chair FuzzyPhoto Project Advisory Committee meetings.
Noted: Alexander von Lunen will be the primary link between the project and partners in order to identify relevant collection segments, establish metadata extraction methods, arrange testing of results and develop design specifications for the end-user interface widget. Jethro Shell will be building the back office databases.

Project Overview

Received: Presentation on the project aims, methods, intended outcomes, schedule and brief introduction to fuzzy logic. Copies of the presentation are circulated with these minutes.
Reported: A project blog has been established at http://fuzzyphoto.edublogs.org/
Action: SB to circulate copies of the project milestones.

How we plan to work together

Tabled: Proposed Terms of Reference for the committee.
Agreed: Membership item 2 to be changed to read “Additional members may be co-opted as appropriate.” Rest of draft accepted. Revised TOR are circulated with these minutes.

Actions: SB to circulate a one page project description that partners can give to potential new partners.
Partners to invite other institutions to consider joining the project. Partners to provide examples or ideas about wider applications of the technology.

Record linkage demonstrations

David Croft demonstrated some of the results of his PhD project so far and explained the methods used.

Online conferencing trials

After some experimentation most members managed to establish a Google Hangout connection.
Agreed: That this is not straightforward and some guidance is required. It will be essential to have a very structured agenda for such meetings which is communicated well in advance.

ACTION: SC to produce a written guide to getting on to the FuzzyPhoto G+ Hangout.

Questions and Answers

1. *In addition to specific object record data, should we include related material such as information about collections or auction house records and catalogues relating to the objects?*

Agreed: While the project will focus on object records additional material will be considered for inclusion and decisions made on a case by case basis.

2. *Do we wish to invite other institutions to join the project?*

Agreed: While the project already has sufficient members to be viable, adding to the number of records overall and to bring in specific collections that are particularly relevant to the content covered by the project would enhance the results.
Agreed: New partner institutions will not automatically be members of the Advisory Committee although they could be invited if there is a compelling case for it.

3. *Should membership of the Advisory Committee be widened to include other specialists, eg. conservators such as Hope Kingsley from the Wilson Collection?*

Agreed: Since the project has budgeted for a specific number of Advisory Committee meetings/members it could place the Committee in an invidious position if it invited some people to join and not others. However there may be a need for a particular skill or perspective and the Committee should be free to invite other people to attend as required.

4. *What metadata standard does the project intend to use?*

Noted: An early project task is to review and determine the most appropriate metadata schema to integrate the different partners' metadata. A strong contender is CIDOC CRM but ICOM's LIDO is also a contender.
Agreed: The decision should be taken in the context of a fuller understanding of the partners collections.
Reported: The V&A are collaborating with the British Museum to align their use of CIDOC CRM.
Reported: SB Is chair of the British Museum CIDOC CRM based Research Space project.
ACTION: AvL to arrange site visits to partner collections at the earliest opportunity to begin identification of relevant content and determination of appropriate metadata. In particular, Alex to look at BM implementation then talk to Heather Craven at the V&A.

5. *How will material be selected for inclusion/What fields is the project most interested in?*

Agreed: To frame the selection initially by the date range of PEIB/ERPS (1839-1915) and to use the date, title and photographer/exhibitor fields as the starting point for comparisons.
Noted: This core set of fields may be expanded.

6. *How often should the Committee members be invited to comment on project activity?*

Agreed: The project team will seek comments on work package reports and on specific issues as they arise. Comments will be collated via google docs and group email.

7. *How should the project be promoted by partners?*

Agreed: Partners to use their existing dissemination channels. It is not necessary for partners to clear project statements with the project team but should alert the project team whenever they make a public statement about the project. In addition Partners can send Alex information to put on the project blog about when they promote the project.
ACTION: Partners to send AvL news and pictures to put on the blog, especially of photographs that are likely to appear in the catalogues.
Noted: Economic impact of the project can be demonstrated by driving traffic to commercial sits eg. SSPL.

8. *Should we use external stakeholders to publicise the project eg. Michael Pritchard/RPS?*

Agreed: All partners should use their existing networks to promote the project as widely as possible.
ACTION: SB to circulate 1 page project summary with contact information for this purpose.

9. *How will the project team access partners' data?*

Agreed: Different solutions will have to be developed for different partners.
Noted: V&A already have an api for accessing their data.
ACTION: AvL to contact partners to discover their data access possibilities.

10. *Will it be possible to engage the wider community in refining the results?, eg Rijksmuseum "windmill group" of amateurs who have authority to tag the records.*

Agreed: Such refinement could be beneficial.
ACTION: HC to ask Windmill Group how people are recruited and how the system works.

11. *Who are the primary audience?*

Researchers, rather than the general public.

FuzzyPhoto Advisory Committee 6 June 2013

V&A, London

Minutes

Present: Marc Boulay, Professor Stephen Brown, Heather Caven, Malcolm Daniel, Thomas Galifot, Dr Simon Coupland, Colin Harding, Professor Roger Taylor
Apologies: Pete James, John Falconer, Dominique de-Font-Reaulx,

In attendance: David Croft, Dr Alexander von Lunen, Jethro Shell, Vicctoria Panter, Susanna Brown.

Introduction

The project team presented an overview of the project covering data acquisition, cleaning, metadata schemas, terminological disambiguation, query expansion, server architecture and security. Examples of record linkage and interface widgets were also demonstrated and a draft bookmark was presented.

Data acquisition

The project now has circa 425,000 records from a variety of sources. The data sets are not always complete, accurate or consistent (eg. one set uses 17 different methods to represent date information). This has required some manual checking and cleaning of data, which will be complete by mid-June 2013.

Metadata schema
After cleaning, the data will be stored in a single data warehouse. This requires a single common data structure. Initially the project envisaged using CIDOC CRM but on closer examination CIDOC is too complex for the project needs and would require the development of one or more suitable ontologies. Given the number of words in each object tile is only around 3 words and yet the subject matter is very broad indeed, developing a sufficiently broad ontology is beyond the resources of the project. Instead the project has adopted the LIDO metadata standard, developed by ICOM for delivering metadata, for use in a variety of online services, from an organization's online collections database to portals of aggregated resources, as well as exposing, sharing and connecting data on the web. The strength of LIDO lies in its ability to support the full range of descriptive information about museum objects.

Noted: LIDO has been adopted by the Europeana Inside project http://www.europeana-inside.eu/home/index.html.

Disambiguation and query expansion

As each object title is only around 3 words, established methods of searching for similarities between different records will not work because the corpus of terms available is not big enough. This problem can be overcome by "query expansion", ie. finding synonyms for the words occurring in object titles. However before synonyms can be found, any ambiguity in the words used in object titles has to be eliminated.
Words used in object titles may have several meanings. For example "fair" in "fair daffodils" may mean blonde, attractive, equitable, carnival. Fuzzy rules based methods are being used to disambiguate the terms available and established thesauri such as Wordnet are being used to expand them.
Noted: Word usage in the 19th century was different, eg. "grotesque" meant elaborate. WordNet does include 19th century terms but weights meaning towards modern terms. However it can be tweaked and trained to favour older terms.

Record linkage demonstration

David Croft showed some results from his PhD study in which he was able to identify co-reference between different objects held in different collections even where there are differences between title, creator, process and date information for each.

Server architecture, security and widgets

Partner data will be held in a single database or "data warehouse" that is physically unconnected to any network infrastructure. Thus remote access to partner data is not possible. The warehouse will be mined for possible similarities between all the records contained and the link information will be exported to another database on a separate server, the links database. This links database will be queried by partner sites whenever a user reaches the level of an object description on those sites. A specially designed widget will

request a list of possible links relating to that specific object from the links database and display the link on the partner Website within the page displaying that particular object. Thus at no time will anyone be able to see the original partner data or even a full list of the identified links between different partner records. All that will be visible at any point in time will be links relating to one specific object with the partner's own Website. Similarly it will not be possible to search the links database directly with a search term. FuzzyPhoto is building a finding aid, not a search engine. It will help users to find objects similar to others they have already found but cannot be used to search *ab initio* across partners sites. Widgets are expected to go live on partner Web sites in July 2014. However partners are not obliged to use Widgets and may wish to customise/develop their own.

Discussion

Widgets

How will the cross referencing work when relating to material on own initial site which is launching the query?
Can the link list for an object prioritise objects from the collections of the institution hosting that particular object?
Can users rate the accuracy of the recommended links?
Can users create their own links as a user if they know a good match already identified?
Can the depth of the search be varied, eg. By a slider to show more or fewer hits? We can show hits in batches of 10 with a "show more" button.
Can we have links at a collection level?
Can we have a process for refreshing records from partners? (NB data structure standards and cleanliness are issues that need to be overcome for this to happen).
Can we group the 'poetic titles' and use their descriptions to search for them instead of their titles?
The project needs to explain the risks to institutions of including widget code.
Institutions need to manage user expectations, eg. If recommended links don't show for all items/other categories.
Need to develop effective ways of drawing attention to links where they exist.
Concern that links to related objects may drive traffic away from institutional sites. Agreed that opening links in a separate window would help to ameliorate this concern.
The project needs to be able to report reliable traffic data for fuzzyphoto.
Need to draw attention to the project and FuzzyPhoto functionality. Can we have some kind of temporary promotional site explaining the project and how to use the resource finding aids. listing project partners, providing usage stats, explaining the project scope, and generally giving weight to it? Could we promote Fuzzyphoto via Alan Griffith's Luminous Lint site?
Proposed to have a rubric preceding the links on each site along the lines of: "Advanced Research. Organisation N has partnered with X, Y, Z to develop an advanced finding Aid that can recommend potential matches with this object. Click on the links suggested to see the related objects." NB this wording can be customised by each site.

Sustainability

The project needs a credible sustainability plan beyond the funded period covering server maintenance, record refreshment. NB the latter has implications for data structure and accuracy.
Can we have batch loader to import test data to allow new partners to test for compatibility? Yes but institutions would have to provide data that conforms to a prescribed the data structure.

Interface design evaluation

The project is to develop set of questions for use by partners in user focus groups to ascertain required design functionality.

Additional partners

The project is open to new partners. National Portrait Gallery has relevant collection. Susanna Brown to provide a contact.

Dissemination

Oracle, the annual gathering of photographic curators (60-100 delegates) is a good venue for dissemination. Marthe Weiss (V&A) is attending in November 2013. Venue in 2014 may be Birmingham and or London.

Bookmark

Include M d'O logo and reverse position of title and strap line.

Actions

1. SB to circulate slides and notes on the meeting.

2. Susanna Brown to provide National Portrait Gallery and Marthe Weiss contacts.

3. SB to provide a clear breakdown of risks to partners about widget installation for each widget option type along with supporting arguments such as increased traffic coming from other partner sites. Gain in traffic is greater than loss!

4. Project representatives to discuss widget implementation with their respective IT teams to obtain approval/agree response. (Partner IT staff to liaise with Simon Coupland for technical discussions).

Appendix 4. Press releases

THES grant announcement http://www.timeshighereducation.co.uk/420843.article

DMU Press release 03/09/12

Date: TBC

Digital detective will save thousands of research hours by tracking down historic photos

Thousands of museums, archives, libraries and galleries will benefit from a De Montfort University (DMU) led international research project which will unearth historical photographic treasures online.

The two-year Arts and Humanities Research Council (AHRC) funded project – named FuzzyPhoto - will recommend potential matches between historical photographic exhibition catalogues and photographs that appear in online collections.

Exhibition catalogue references are a printed list of exhibits at an art exhibition. Early exhibition catalogues in the nineteenth and twentieth centuries often had no pictures, relying on written descriptions of the photograph on display. This makes assigning a reference to a specific photograph a complex process involving considerable time and travel for researchers to match relevant material.

FuzzyPhoto will automate the process, developing 'finding aids' that will be able to recommend potential matches between historical exhibition catalogue entries and images of photographs in online collections even where there is not a precise match.

The £390,000 project will be hosted by DMU's Photographic History Research Centre, in collaboration with the Centre for Computational Intelligence.

Professor Stephen Brown, Head of DMU's School of Media and Communication, explained: "Image collections are increasingly being published online and search engines are becoming increasingly powerful, creating a timely opportunity to match photographs with their original exhibition catalogue entries without travel to numerous archives around the world.

"Within the UK alone over 1,000 museums, archives, libraries and galleries could benefit from this research, enhancing the value and utility of their collections and of their online services through increased information, improved accuracy and functionality. Commercial dealers and auction houses will find it useful for attribution and value, while more accurate identification of artefacts such as photographs can help prevent the inadvertent export of nationally important treasures."

The project starts in November 2012. Global partners are Birmingham Central Library; the British Library; the National Media Museum at St Andrews University; the V&A; the Musee D'Dorsay and the Louvre in Paris, the International Council of Museums and the Metropolitan Museum in New York.

Ends

Notes to editors

The FuzzyPhoto project derives its name from investigating the potential of combining probabilistic record linkage with 'fuzzy clustering' to identify co-reference between exhibit records and published images.

DMU is a university of quality and distinctiveness in the heart of Leicester. We are distinguished by our life-changing research, dynamic international partnerships, vibrant links with business and our commitment to excellence in learning, teaching and the student experience. We celebrate the rich cultural diversity of our staff, students and all our partnerships.

For more information please contact the De Montfort University Media Office on 0116 2577674 or news@dmu.ac.uk

Follow us on Twitter at www.twitter.com/@dmuleicester

Appendix 5. Project bookmark

Appendix 6. Partner visit reports

Report on the visit to the National Media Museum, 30/11/2012

Alexander von Lünen

AvL visited Colin Harding on 30/11/2012 to discuss the nature and structure of the data to be delivered by the NMeM. The topics discussed were:

Fields

AvL presented CH the list of fields the FuzzyPhoto had devised on their meeting on 13/11/2012. CH agreed to this and couldn't think of any other fields required in order to make the FuzzyPhoto project work.

Software

The NMeM uses two database systems: Mimsy XG and iBase Manager 9.45.1; long-term strategy is to abandon iBase in favour of Mimsy, but since record management is principally done by the Science Museum Group (of which the NMeM is part of), this is out of the control of the NMeM. iBase Manager is no longer sold to new clients, according to the producer's webpage (www.ibase.com), so a switch-over seems to be inevitable. Right now, only a subset of the records in iBase is in the Mimsy based system, and all of the photos in the Mimsy system should be available via http://collectionsonline.nmsi.org.uk. However, a query revealed that there are only 1762 photos available on that website (advance search for NMeM photos between 1853 and 1915), whereas the iBase database has 58,945 records in it.

Data transfer

As usual, this has to go through the right channels, with managers at NMeM (and possibly the Science Museum Group) giving their ok. CH will pursue this with the right people. Once this is sorted, a data transfer can be done almost instantly. A transfer before Christmas looks feasibly, provided the MoU has been signed by one of the NMeM managers by then.
AvL and CH consulted with Pete … (the web manager) about the technical issues in the data transfer. He's happy to help, and suggested a one-off data dump. Everyone present agreed that it might be an idea to have a preliminary data dump ASAP, and another (perhaps more voluminous) one closer to the project's deadline. Mimsy XG allows data export as CSV, TAB and XML, so there shouldn't be any problems in terms of format. AvL couldn't find much technical information about export formats in iBase manager on their webpage, but it mentions a REST interface for the web version (and that the V&A is using it).

Other

AvL and CH discussed other issues as well. CH suggested it would be a good idea to have other information/links included in the search results as well, not only links to collection records, such as related webpages (e.g. the entries in www.scienceandsociey.co.uk or www.ingenious.org.uk, or – for example – CH's video on Fenton's still photography on the NMeM website could be linked to when searches on Fenton and still lives are searched for. AvL agreed that this would be desirable, but that it's likely to be out of the scope and resources of the FuzzyPhoto project.
Something more desirable, obviously, would be a link between the two databases at the NMeM (i.e. the Mimsy based and the iBase based systems). As mentioned, the Mimsy based systems only holds a subset of the iBase version and migration from the latter to the former may take a while at NMeM. It would be nice to expose the records of the NMeM by creating links between the two databases, but again this will be likely beyond the scope of the FuzzyPhoto project.
In regard to photographic material to be used for the FuzzyPhoto flyer/bookmark, CH suggested to ask Kelly Wilder, as she should have received images from the NMeM for an MA brochure some time ago. We are free to re-use those. Pete will send AvL a high-res logo for the NMeM to go on the flyer (so far, AvL used the logo from the NMeM website for his prototype; the logo on the website being outdated, above all).

Actions

1. CH to pursue that the MoU is signed by the responsible person at NMeM;

2. Following from that, AvL liaising with Pete to have a data dump transferred, preferably in XML format; these two points should be achieved before Christmas;

3. Pete to send AvL a high-res logo of the NMeM to go on the FuzzyPhoto flyer/bookmark, asap;

4. AvL to get in touch with Kelly Wilder to get NMeM photos that she received previously, asap (maybe to be addressed in person during the drinks meeting on 04/12/2012 in the department);

5. AvL to look into how to incorporate the URLs from the Collections Online website into the database. There are no links in the NMeM Mimsy system, i.e. there's no information in the Mimsy based records manager where to find a particular image online.

Report on the visit to British Library, 17 Dec 2012

Alexander von Lünen

AvL visited John Falconer at the British Library on Monday, 17/12/2012 to discuss the BL's contribution to the project. The topics discussed were:

Fields & BL Record Management System

The BL Photography section uses a custom record management system (RMS), a web application based on MS-SQL. A complete version is only available internally to the BL, but a subset is used for the catalogue on the BL website: http://www.bl.uk/catalogues/photographs/
While the RMS doesn't follow any particular standard (such as SPEKTRUM), the field structure seems to be quite common, i.e. the structure is not so different from other systems we've encountered so far. The fields are pretty much what we want (e.g. size, process, date range, etc), and would comply to standards like LIDO quite readily.
The identifiers for the photos are defined by BL and follow the form Collection/Part (Folio/Print); for example: Photo20/1(25), refers to the collection "Photo20", "Part 1" (a collection might be broken up in several parts), and folio/print 25 within this part of the collection. "Collection" in this context refers to the IDs given by the BL, i.e. there's no collection "Royal Photographic Society" as collection identifier (this would have to be full-text searched for in the description field). The collection names and other terms used for the catalogue entries are stored in an authority file by the BL.
None of the photos is online, because a) of time/budget issues and b) copyright issues. The current photographic database of the BL, which is quite specific and unique, is supposed to be migrated into the general BL catalogue. This may result in some data being purged from the RMS to make the data comply with the library catalogue structure. However, no specific date has yet been set for this.
Oddly, the system could give no figure of how many records it holds.

Data exchange

Data exchange and the MoU will be discussed at our meeting in January. JF anticipates no problems. They have exported data into a CSV file before, so this can be done. Since the data model is non-standard, the BL database person will have to craft a query to get all the data we'd want collated and then export it. It might be worth considering getting the whole RMS db as dump, putting it into our own MS-SQL Server, and then deciding what data we would want. JF will send AvL a list with the fields present in the RMS frontend.

Misc.

JF is happy to send a high-res version of the BL logo and possibly some photos for us to use for the bookmark, once the MoU has been signed.

Actions

1. JF to send AvL the list of fields of the RMS; AvL to remind him if necessary.
2. AvL to analyse the BL publicly accessible online photo catalogue to assess what kind of information we can expect.
3. All other activities depend on the meeting on Jan 21, where the MoU and technical details will be discussed.

Report on the visit to Birmingham City Library, 18 Dec 2012

Stephen Brown

Stephen Brown, Alex von Lunen and Jethro Shell visited Pete James and Rachel McGregor at the Birmingham City Library on Tuesday, 18/12/2012 to discuss the Birmingham's contribution to the project. The topics discussed were:

Library closure schedule

Public access to the photographic collection has already closed, in preparation for the move to the new library. The new library will open on September 3rd 2013. Physical access to the photographic collections will be restored by October/November 2013. Online catalogue access will be maintained throughout. (Rachel to confirm this).

Library Record Management System

The photographic collection is only partially catalogued. Some of the catalogue records reside in card index form only. Digitised records are accessible via an online interface to a CALM database (http://www.axiell.com/calm). The database is maintained by a third party supplier Service Birmingham Ltd. None of the photographs are digitised.

Selection of records

The Birmingham Photographic Society (BPS) collection is the most suitable set of records to begin with since it aligns with the project data range (1839-1915) and is fully catalogued online. Further collections may be selected later in the project.

Memorandum of Understanding

SB explained the intentions behind the proposed MoU and stressed that (a) the precise wording can be changed providing the meaning is mot altered and (b) it is not necessary for the MoU to be signed before data transfer begins but the MoU does need to be signed before the project ends.

Library contribution

The estimated 15 days contribution of Library time was reviewed. SB emphasised that it was an estimate only, that partners would not be held to the figure of 15 days and that the structure of the project allowed the contributions of different partners to be flexible. Because of the closure and relocation of Birmingham Library it was recognised that beyond providing access to records at the start of the project Birmingham's contribution was likely to be weighted towards the back end of the project, particularly in relation to public dissemination activities.

Miscellaneous

Tom Epps and Rebecca Cadwallader are working on crowd sourcing ideas that may be relevant to the project later on.

Actions

1. RMcG to confirm continuity of availability of CALMview.
2. AvL to analyse the publicly accessible BPS records to assess what kind of information we can expect.
3. PJ to investigate access for AvL to download a dump of the BPS records.

4. PJ to refer MoU text to Library senior managers/legal representatives.

Report on the visit to the Musée d'Orsay and the Bibliothèque Nationale de France, 11 Jan 2013

Alexander von Lünen

Musée d'Orsay (MO)

Present:

1. Stephen Brown (SB), DMU
2. Alexander von Lünen (AvL), DMU
3. Thomas Galifot (TG), Curator for Photography, MO, thomas.galifot@musee-orsay.fr
4. Dominique de Font-Reaulx (DdFR), Curator, Louvre
5. (later) Françoise Le Coz (FLC), IT officer, MO, francoise.lecoz@musee-orsay.fr

TG is very interested in the project and happy to cooperate. The MO collection holds c. 46.000 objects. The entire catalogue is online (http://www.musee-orsay.fr/en/collections/index-of-works/home.html?cHash=1030a57d48). The data is stored in an Oracle database and might get migrated to a newer version in a year's time. We could grab the records from the web catalogue, but FLC is happy to give as an export as Excel-Sheet (maybe other formats as well, such as an SQL dump, AvL to follow up). In addition to that, MO is participating in a museum/collections portal: http://www.photo-arago.fr. This, however, only holds a few selected photos/records from the collections and might not be of too much use for our project.
Judging from the web catalogue, the digital records are quite complete and well documented. All those present agreed that it would make more sense for MO to give us the entire records and have DMU decide what might be useful for the project, rather than MO filtering the records by, say, time range (e.g. C19) or geography (e.g. exhibitions in Britain only).
Actions:

1. TG to get ok from administrators to hand over the records data; and
2. AvL to liaise with FLC once this ok has been given to discuss the data transfer.

Bibliothèque Nationale de France (BNF)

Present:

1. SB
2. AvL
3. DdFR
4. Sylvie Aubenas (SA), Curator for Photography/Head of Department, BnF, sylvie.aubenas@bnf.fr
5. [didn't catch her name], IT officer, BnF

SA very interested in the project, quite curious about search engine algorithms and fuzzy matching; would love to do something with image recognition algorithms. The BnF has a huge collection of objects (TG from the MO used to the phrase "only 46.000 objects" for their collection to relate to the size of the BnF collection); they use different systems and catalogues to store the records, simply because of the size and the age of the collections. They will get as much information out to us as they can; the IT officer mentioned that they can give us an export in OAI (Open Archives Initiative) format, which would be ideal.

Actions:

1. SA will contact the BnF directors to discuss data exchange issues and have the MoU signed;

2. AvL to liaise with IT officer to get the data.

Overall impression

Both MO and BnF seemed very interested in the project, especially in the aspect of networking the data across institutions. The systems (i.e. the digital records managements systems) seemed to be well organized and accessible. Particularly, there doesn't seem to be the situation (as with many UK institutions) of having records scattered across different record management systems, because of legacy and resources issues.

Report on the visit to the V&A, 15 Jan 2013

Alexander von Lünen
Present:

1. Alexander von Lünen, DMU

2. Stephen Brown, DMU

3. Heather Caven, V&A, Head of Collections Management and Resource Planning

4. Susanna Brown, V&A, Curator Photographs

5. Victoria Panter, V&A, Documentation Manager

6. (left c. 13:00) Glenn Brown, V&A, Head of Research

We had a long and very productive meeting at the V&A; the quality of their data and their collection management system (MuseumIndex+) are impressive. The V&A have c. 500.000 photographs in their collection; 3-400.000 are in the CMS. It was agreed that it makes more sense if the V&A let us have all of the records' metadata, rather than filtering it by, say, date. V&A is happy with giving us all the data (the MoU had been signed before Christmas).
Victoria will extract a sample data set of 200-300 records and send it to DMU next week. MuseumIndex+ can export to CSV and XML. Modes of data transfer will have to be discussed when it comes to transferring the whole batch, as this file will be fairly large. The V&A also has a Web API where records can be downloaded, but it has size limits, i.e. one cannot download an infinite number of records. It was agreed that it was much more tenable to transfer the data by other means.
The fields DMU would like to have have been reviewed and discussed. The CMS at the V&A is very granulated and detailed. A lot of the fielded out metadata would be quite interesting to have, but since none of the other partners have records at that detailed level, it wouldn't make much sense to have it from the V&A. Some fields from the V&A CMS may have to be concatenated to fit into our data model; on some occasions, on the other hand, the records consist of not much more than a (rather meaningless) title and a short description. V&A will do a data analysis to get a figure of records with sparse and rich data to find out whether there might be an issue. They will also give us a copy of their cataloguing standard, so DMU gets a better understanding of the V&A data.
Lastly, the V&A would be happy to host future committee meetings. Their location in London might make it more convenient for the other partners to travel, and other members of partner institutions (such as Victoria and Susanna) would have a chance to join the meeting.

Actions:

1. AvL to send Victoria Panter an e-mail with the desired fields and a brief explanation of them ASAP.

2. Victoria Panter to send AvL a sample data set of the V&A CMS in week 4.

3. V&A (Heather/Victoria?) to send catalogue standard to DMU.

4. V&A to do data analysis of the CMS records to get an understanding of distribution rch vs sparse records.
5. AvL to think about (and discuss with Victoria) which export format would be best (CSV vs XML).
6. AvL and Victoria to discuss how the final data dump will be transferred to DMU.

Report on the visit to British Library, 21 January 2013

Stephen Brown
SB visited John Falconer and Adam Farquhar at the British Library on Monday, 21/01/13 to discuss the BL's contribution to the project. The topics discussed were:

FuzzyPhoto project

SB gave a presentation on the project background, goals, methods and schedule and we discussed the BL contribution for each work package.

Content selection and data exchange

The BL Photography section owns 250,000 records. It was agreed that the BL will supply DMU with all of these records, probably as a CSV file in the first instance. The BL National Bibliographic Archive has a SPARQL end point and the BL is willing to explore the possibility of setting up a similar access point for the photographic catalogue data later in the project. The contact for access to the data is Simon Woolf [simon.woolf@bl.uk].

User interface design

The BL has a user interface design specialist who should be involved in determining the specification for the BL user Web interface.

Memorandum of Understanding

The MoU was discussed and agreed in principle. Agreed that the project should proceed pro tem while BL contract specialists review the wording, aiming for signature by the end of Summer 2013.

Images

JF is happy to send some photos for us to use for the bookmark.

Actions

1. SB to send AF a copy of the MoU.
2. JF to send AvL some sample pictures.
3. AvL to contact Simon Woolf to request a copy of the catalogue records. It is important to cc John Falconer into this request.

Report on the St. Andrews Visit, 18/02/2013

Alexander von Lünen

Present:

1 A.v.L., DMU

2 Marc Boulay, St. Andrews University Library

The system

The record management system used is KE-Emu, a product by -- you guessed it -- an Australian company. The system uses a hierarchical data model; on top is a collection, divided into n series, which holds m images, for

which there can be x objects; i.e. the term “image” refers to a picture being taken, the “objects” are then different instances (e.g. prints) of that image.
The nice thing is that the system also holds descriptions for collections and photographers. The system has c. 120,000 image records, out of which 67,000 entries are “clean” according to MB, i.e. the records contain enough and consistent data. The easiest way to do an export would be in CSV format. The system can store export profiles, AvL and MB have defined one for the FuzzyPhoto project, meaning that an export can be easily done. MB is thus happy to deliver an update of the data at any given time later in the project.

The data

Data sets seem to be rather consistent and rich. Dimensions are often not given, or are part of the description field. Dates for collections are usually the creation dates, dates for objects are usually the printing dates. In c. one-third of the records, no title or description infos are given; however, there is almost always a number of subject terms available, so the records would still be useful to a degree. The system uses the term “Originator”, which in 90% of the cases (according to MB) means the photographer. There are no URLs in the system to the online catalog. But they can be easily generated by using the Object Number from the records, and copying&pasting that into a query URL from the website.
With all filters applied (such as date range 1835--1930, or sufficiently populated fields) we found that we would get 20,041 records from St. Andrew, although this might change, as the RMS is still being worked on.
The system also holds the biographical information for every photographer, which in theory we could also get. However, this would require some legal legwork. The biographical data was collected by Sara Stevenson, formerly National Portrait Gallery. There would be the issue of who holds the copyright in that (Sara or the NPG), so if we wanted it, we or MB would need to talk to Sara and sort this out.
The system also holds data for “Parties”, i.e. the persons depicted in an image; usually short personal info, no full-fledged biographies.

Actions

- MB to send data the week after the visit
- We can have further updates any time we want
- We can have images any time we want -- AvL to ask MB for them asap

Appendix 7. WP 3 Data Ingestion and Warehouse report

FuzzyPhoto

AHRC AH/J004367/1

Abstract

FuzzyPhoto is a two year AHRC funded research project that is developing and testing computer-based "finding aids" that can recommend potential matches between historical photographic exhibition catalogue records and images of photographs that appear in online collections, even where there is not a precise match. Work package 3 covers the work required to import the meta-data from the project partners into a unified data model from which the links can be harvested.

Distribution Type: Internal

Author: Dr Alexander von Lünen

Keywords: FuzzyPhoto, metadata, historic photographic collections, museums, libraries, catalogue data, database technology, data import procedures, LIDO, data warehouse, data ingestion

Contents

Summary

This document details the process of importing the project partners' data into the (temporary) tables of the FuzzyPhoto database. Data was usually delivered as CSV or XML files, and each of these formats required different processes. Each dataset is quite unique, so it has not been possible to adopt a unified approach to importing the data. In the future, it would be advisable to ask the partners for a standardized data format.

The individual datasets were first imported into MySQL as specific tables, after which some data cleaning was carried out and then the data was transferred into the chosen metadata schema (LIDO), where some more cleaning was conducted, (see below).

Status

Outputs of this workpackage comprise:

1. This report
2. MySQL temporary tables
3. MySQL LIDO data warehouse comprising:

Partner	Records before cleaning	Records after cleaning
Birmingham City Library	5,513	5,455
British Library	28,974	28,925
CultureGrid	172,148	171,840
ERPS	34,197	34,197
Metropolitan Museum	9,526	9,526
PEIB	20,453	20,453
Musee d'Orsay	46,229	46,228
National Media Museum	8,380	8,380
National Museums Scotland	14,915	14,883
St Andrews University Library	18,620	18,604
Library of Congress	875,267	875,267
Brooklyn Museum	2,352	2,352
National Archives	73,187	71,958
Victoria and Albert Museum	101,538	98,598
		1,406,666

The elapsed time for this work package was 9 months. The resources required to complete this work package were 70 person-days (partner liaison, etc. not included).

Objectives

To import the catalogue data from the project partners – which was delivered in diverse formats – into first a temporary relational database and then into a unified relational database based on the LIDO schema.[3]

The data from the project partners turned out to be quite heterogeneous and required a good deal of work to unify the various data sets into one schema, necessary to generate the links between the different data. The heterogeneity consists of very different structures of the partners' data, usually introduced by the respective record management software (RMS). Each project partner uses a different RMS with varying export capabilities, limiting the extent to which exporting the data can be customized and therefore made compliant to a specific data model. The decision to run a specific RMS is often driven by legacy issues, i.e. availability of a certain RMS package at a certain point which was then populated with the record data. Switching to a different RMS would be too costly for partners and is thus not an option. This meant that few of the data sets could be tailored to comply to a given schema. Therefore, rather than loading the data directly into the chosen metadata (LIDO), temporary tables were created to collate the data and run “clean up” scripts on it.

It was originally considered to develop a batch loader to achieve this, but delays in the data submission by some partners and the very diverse nature of the data made this enterprise not very feasible. The process of importing and cleaning up the data, on the other hand, yielded a very good impression of what would be involved in creating such a tool.

[3] http://www.lido-schema.org

The metadata schema

It was first considered to use CIDOC-CRM as an ontology-based data model for the database. However this turned out to be more a hindrance to the project rather than a benefit. CRM is very complex and hard to keep minimal, i.e. there are a lot of mandatory data fields whereas there are very few with LIDO. Given that the meta-data delivered by the partners was so diverse, it would have been hard (if not impossible) to comply with CRM, and would have been necessary to edit the partner's data to a greater extent (see below for a discussion LIDO vs CIDOC CRM).

The complexity of CRM can also be a bit overwhelming, with some object types almost getting into philosophical realms. Since the fuzzy logic software developed by the project does not take the extra features offered by an ontology into account (such as a dedicated reasoner), it wouldn't have been very feasible to use CRM. It was therefore decided that LIDO offers the best compromise in terms of expressivity, flexibility and implementability.

LIDO was devised in 2010 purely as an exchange format to transfer data independently from the RMS being used. Several major players were involved in its design, such as CDWA Lite, Athena, CIDOC-CRM, Spectrum (CollectionsTrust, UK), MuseumDat (Germany). It has established itself as a standard in the GLAM-community since its inception. One of its major benefits is its versatility, i.e. that it allows flexibility without sacrificing expressivity.

LIDO is an event-based data model, i.e. its main concept rests on the “things” that can be “done” to an object, such as creation, acquisition, restoration, exhibition, etc. This allows for very detailed metadata to be stored and the life cycles of an object meticulously recorded.

“Harvesting Formats” vs Ontologies

A line must be drawn between so-called “harvesting formats” (HF) and full-fledged ontologies. HF in general are formats for publishing and interchanging data, independent from a specific database software etc. Ontologies, on the other hand, also describe the data so being published, i.e. ontologies operate at the conceptual level. Simply put, HF and ontologies differ in regard to the amount and level of metadata they incorporate. HF, as mentioned, is about publishing collection records, so metadata usually concerns things like data of creation and creator of an object, for example. Since ontologies were devised in the context of Knowledge Representation Systems, they must provide enough metadata for automated reasoning. For example, the LIDO format, defines a field for the gender of an actor (such as creator or owner of an artefact). For an HF it is good enough to have different identifiers (such as male, female, unknown) to operate, whereas in an ontology there must also be rules about the “nature” of these different genders. “Nature” here referring to a set of logical constraints; for example, that “male” and “female” are mutually exclusive, i.e. when an actor is labelled as being “male”, the “reasoner” (a piece of software that evaluates the ontology) can infer that the actor is not “female”. This may seem trivial to humans, but in an ontology rules such as these need to be defined to allow the reasoner to work efficiently; the point here being that the strength of ontologies is to allow transitive queries. If, for example, I would query for female artists in the database, but not all of the entries have their gender field set, the only fix would be to update the gender field before I could run the query. In an ontology, however, I could define a set of rules that would let the reasoner infer that the record being looked at must be a female person, e.g. by having occupation “actress” and there being a rule in the ontology that an occupation of type “actress” refers to a female person.

So, to summarize, GLAMs might quite likely have data already that could be described as “harvesting format”, usually the collection records. But it is perhaps no so likely that they have a full-blown ontology for their collection. Luckily, standards exist for both (HF and ontologies), and the the following sections will discuss the best candidates and how they could be utilized in the FuzyPhoto project.

Harvesting Formats

There exists a number of HF; the main reason being that metadata about collection items have been recorded for a long time now, starting with paper cards in filing cabinets, and XML has been around for some time now acting as lingua franca of structured data exchange format.

Many different HF have been designed in the past years, usually along national and topical interests (i.e. depending on the type of collection). There seems to be a clear trend, however, to find an internationally agreeable standard. This has resulted in the LIDO (Lightweight Information Describing Objects, `http://www.lido-schema.org`) XML schema. Other, more national, schemas such as Museumdat by the German Museums Association (`http://www.museumdat.org`) have voiced their support for LIDO. It thus seems to be a promising candidate for getting an exchange format for GLAM data.

LIDO only defines a minimum set of mandatory data fields to be populated, making it rather flexible. Given that is has been developed by several museum organizations, it should be quite close to the kind of data that is frequently encountered in the GLAM sector.

Ontologies

The selection of ontologies in the GLAM sector is much more limited. For years Dublin Core had been the least common denominator, with various national bodies (LOC, DNB, etc) rolling their own standard. As of late, CIDOC CRM strives to be the gold standard in the cultural heritage sector. While quite complex, it has been closely modelled after the collection management philosophies in the GLAM area and is experiencing ever more support from it. CIDOC CRM offers a multitude of classes to describe collections and related concepts. It thus makes a good candidate for storing the project data on the search engine side.

Strategy for the project

That being said, it should be obvious that many of the project partners in the GLAM sector may very well have data in their respective proprietary collection managements systems that could be easily mapped (i.e. exported/converted) to LIDO, while it is rather unlikely that they have sufficient metadata for CIDOC CRM. There are, however, tools and manuals on the CIDOC website to transfer LIDO (or general XML data) to CIDOC CRM, so the door is not closed on that one.

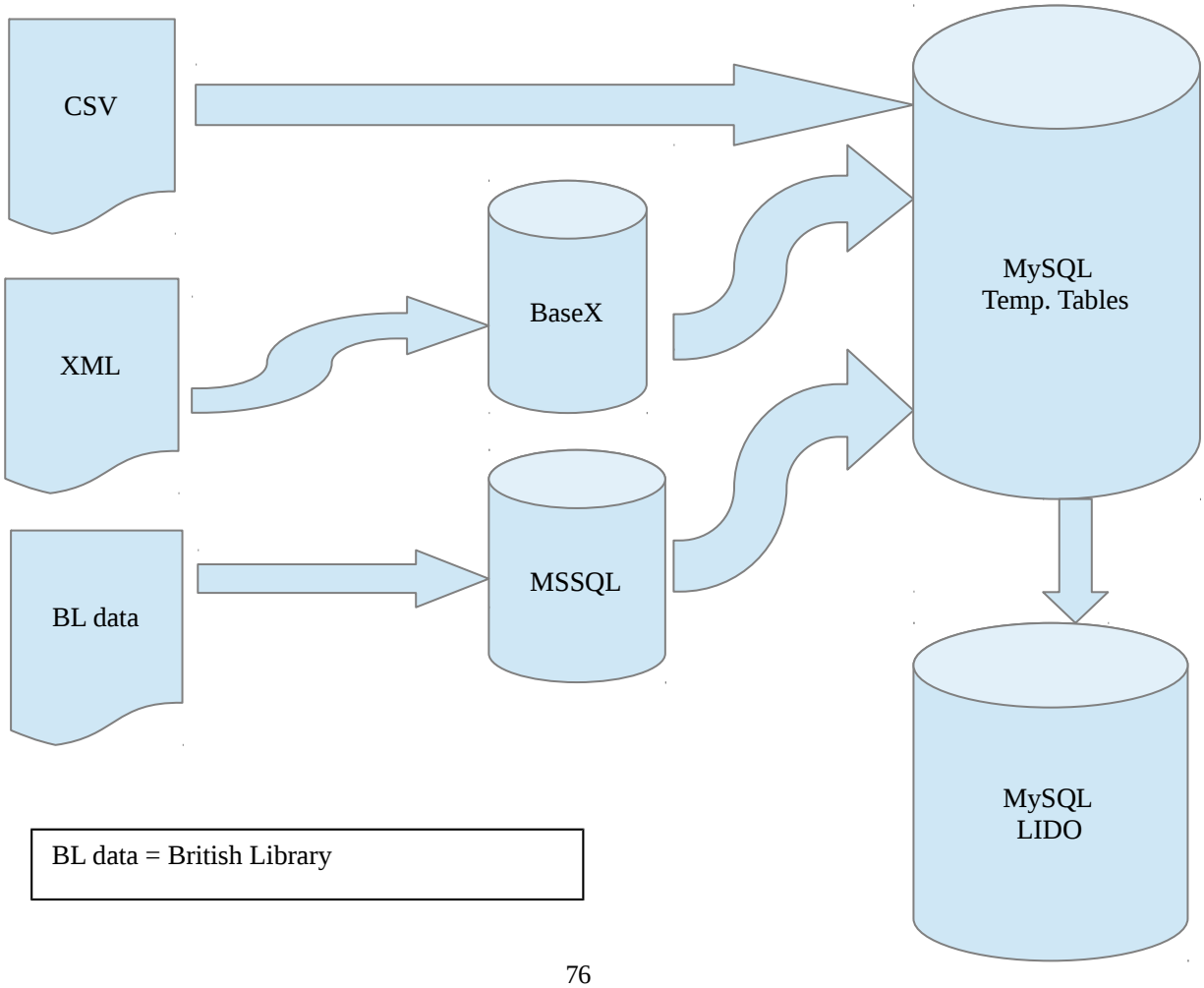
It thus makes sense to agree that the LIDO format is used as an interchange format within the FuzzyPhoto team, i.e. the project partners will deliver their data in some kind of structured data (possibly LIDO, but other formats are permissible) to the team, which will convert it to the LIDO format in order to unify the different data sets.

Requirements

As the data from the partners' is very diverse, a range of software tools were required. For XML data, an XML data store and XQuery queries were readily utilized (see below). For tabular data (CSV), the import facilities of MySQL usually sufficed to handle it. After the data had been imported, “cleaning up” the data – such as eliminating duplicates or spelling variants – was done chiefly with SQL queries with some user-defined functions (UDF) found on the Internet. Some operations, however, were too complex to be handled in SQL, so Java and Python scripts were used externally.

The software dependencies were thus:

- BaseX: as XML data store
- MySQL: as database, both for temporary tables and the LIDO schema
- LibreOffice Base: to aid in importing some of the data
- LibreOffice Calc: dto.
- MS SQL Server Express 2008: for the BL data
- MS SQL Server Express 2008 client library: for transferring BL data to MySQL
- MySQL Workbench for Windows: dto.
- Python
- Java



General Issues

The schema name “fuzzyphoto” was used to hold the partners’ data in the MySQL database. This schema served as temporary data store from which data could be transferred into the LIDO schema, which is held in the schema named “mydb” (this name was assigned by the ERM modeller that comes with the MySQL workbench and that was used to model the LIDO XML schema into a relational model).

The data from the partners had some issues that made it hard to clearly distinguish events and actors etc. For instance, in many data sets it isn't always obvious whether the date specified refers to the date of the shoot or the date of the print (and maybe this is difficult to establish from the source). Since “Shoot” and “Print” are two events in the LIDO model, it would be relevant to know the difference, while it should be ok to neglect this for the fuzzy logic algorithms.

Another obstacle was that field names in the partner's data were not consistent, both across the data sets and at times even within the same data set. For example, the sample data the V&A provided had different field names than the final data set we received. It would be advisable to prescribe for the partners the names of the fields they have to use in their data set so that a batch loader can just pick them up, rather than having to write a configuration file that labels fields correctly.

Importing CSV files

Importing CSV into MySQL is very straight-forward, using the load data command within the `mysql` shell. However, the table the data is imported into needs to be created first. For example, to load the data from the National Museum of Scotland one would create the table in `mysql` first:

```
create table nms (id varchar(100), description text, type
varchar(100));
```

Then, the command to load the data would be issued (make sure to get the field delimiters right, these are just the ones I have been using):

```
load data local infile 'nms.csv' into table nms fields terminated
by '\t' enclosed by '”' lines terminated by '\n' (id, description,
type);
```

The order of the columns need to be specified in the order they appear in the CSV file. Also, types need to be specified correctly in the create statement, or the data won't be imported. It is advisable to create all columns as string types (i.e. either varchar or text), and convert them them afterwards, as typos etc. might prevent import. (Note: in many cases I used the vertical bar character “|” as field separator, as this is far less ambiguous).

NB there is also a command line tool named `mysqlimport`, which does pretty much the same job, but you don't have to log in to the `mysql` shell. The syntax for that would be:

```
mysqlimport --fields-terminated-by=”\t” --lines-terminated-by=”\n”
-p -u {username} -v -L {database_name} {filename}
```

One would still need to create the tables first, though, so using the mysql shell might be handier. On the other hand, mysqlimport is much faster than the load data command, so for large data files this

will be much more efficient. Note that there is no option for mysqlimport to specify the name of the table in the database the data is supposed to be loaded into. The tool assumes the file name is the same as the table name, and your only option would be to rename either the file name or the table's name within the database.

Importing XML files

Importing XML proved to be a lot more complicated, as the schemas were quite diverse and each data set required a different approach. After trying XSLT scripts, it was decided to use a XML store, BaseX (`http://www.basex.org`), to do the transfer. The main reason is that BaseX provides a GUI and visualizations, making rapid prototyping much easier.

As a first step the XML files are loaded into BaseX. BaseX has the option to do bulk imports, so that when the data comes spliced up in several files they can be easily merged within BaseX. The software also offers different visualizations, such as a tree view of the XML data, which helps analyzing it.

BaseX has an XQuery shell integrated. After some experimenting, it was decided to create individual XQuery queries for every data set that would generate SQL Insert statements, which then can be run inside MySQL (there are XQuery software libraries that allow JDBC connections, so that the data could be exported to the MySQL database directly; however, these libraries were only available as commercial versions).

In terms of modelling the XML data as relational database tables, every nested element (i.e. sub-branch of the higher-up nodes) in the XML document tree were modelled as their own entity/table, linked by the object Id of the top node, if the analysis showed that they could have more than one entry.

For example, the V&A data has – among others – an object for dimensions:

```
<dimensions>
   <part name="(image only)">
      <dimension field="Height" unit="cm">40.9</dimension>
      <dimension field="Width" unit="cm">27</dimension>
   </part>
   <part name="(mount)">
      <dimension field="Height" unit="cm">47.5</dimension>
      <dimension field="Width" unit="cm">35</dimension>
   </part>
</dimensions>
```

The SQL create statements would look like this (this, obviously needs to be done prior to the import):

```
CREATE TABLE vanda_dimensions (
  id varchar(50) NOT NULL,
```

```
  name varchar(50) DEFAULT NULL,
  width varchar(255) DEFAULT NULL,
  height varchar(255) DEFAULT NULL
);
```

"Id" refers to the object id, "name" is the type of dimensions (i.e. mount/image/etc.) and "width" and "height" are the actual dimensions. The XQuery script would thus produce an `insert` statement like this:

```
INSERT IGNORE INTO vanda_dimensions (id, name, width, height)
VALUES( 'O10003' ,  '(image only)' ,  '27cm' ,  '40.9cm' );
```

```
INSERT IGNORE INTO vanda_dimensions (id, name, width, height)
VALUES( 'O10003' ,  '(mount)' ,  '35cm' ,  '47.5cm' );
```

NB the brackets around "mount" etc. are removed by the SQL clean up scripts.

Cleaning the data

Cleaning the data is done within SQL once the data has been imported. In some cases it made more sense to do this once all the individual datasets had been transferred into the LIDO schema, as some issues occurred in more than one dataset (such as date formats). The cleaning operations for the individual data sets will be described in the next sections, the one for the general LIDO schema is detailed further below in the section on the LIDO implementation.

Individual Datasets

Birmingham City Library

The data from BCL was received as Excel sheet. Exporting from LibreOffice Calc to CSV and importing it into MySQL was no challenge.

The data needed some clean up (see `clean_birmingham.sql`), chiefly to correct typos, get spelling consistent and to unify date formats.

British Library

The dataset from the BL proved to be the most difficult to import, simply because we received it as MS SQL Server 2000 dump file. There was no MS SQL Server instance running on the computers the team had access to. Since installing it on one of the other Windows machines required Admin rights, AvL used his private laptop for this. The process was very messy and required a lot of trial and error, as the migration wizard would make wrong assumptions and data types etc. had to be adjusted manually before it succeeded. It is highly recommended to ask the BL for a different format in the future. As pointed out below, MS SQL Server is used by the BL as XML data store, so exporting the data as XML file (as done by AvL, see below) should be feasible.

The process ran as follows (all on the MS Windows machine):

- Download and install SQL Server Express 2008 (from `http://msdn.microsoft.com`). One needs to search a bit for this version, as the latest version (SQL Server Express 2012) is not downward compatible enough for a SQL Server 2000 dump file.
- Download and install the SQL Server 2008 Client library (for ODBC).
- Install MySQL Workbench v5.2.4+ (for the migration wizard).
- Configure MySQL server on the Linux box which holds the FuzzyPhoto db to allow connections on port 3306:
    - `iptables -A INPUT -i eth0 -p tcp --destination-port 3306 -j ACCEPT`
    - in `/etc/mysql/my.cnf` change/add this line: `bind=0.0.0.0 !`
    - `reboot`

5. Restore the dump file received from the BL (in this case it was named `JerwoodDiscovery.bak`) into SQL Server Express 2008 (can be done with the GUI tools that come with the installation package).
6. Configure an ODBC DSN in your system settings (here it was given the name “mssql”). Important: Use path for server, i.e. `AVL-PC\MSSMLBIZ`
7. Use the data migration wizard in the MySQL workbench to transfer the data from the SQL Server 2008 Express database (via ODBC) to the MySQL database running on the Linux box.
    - There was an error during this process coming from the schema DDL when choosing automatic migration: “Wrong syntax for column timestamp in table TcRecords”. The reason being that the SQL Server 2000 data type timestamp is different from the

SQL-92 standard timestamp data type.

- To fix this, do the following:
    a) click “Back” twice from this point to get to “Manual Editing”
    b) choose “All Objects”
    c) pick table *TcRecords*, go to columns
    d) change the DDL from `timestamp TIMESTAMP(0) NULL` to `timestamp BIGINT DEFAULT 0`
    e) in “Manual Editing” delete the index `UNIQUE INDEX IX_Key …` from the DDL; this may raise an error, which you can ignore, just re-create the index in the MySQL database after migration manually

The data should now be in the MySQL database on the Linux box. There are quite a number of tables, most of which seem to be redundant or otherwise unnecessary. It turned out that some of the tables held straight tabular data, while others were used to hold much of the same data in XML format (i.e. SQL Server 2000 is used as XML store by the BL). After some analysis and a query to the BL, it was assessed that most of the XML data was the more recent and complete version. Thus, it was decided that it is easier to dump the XML data as a text file from the db and then re-import it as SQL data in the way described above (“Importing XML data”), rather than trying to do this within MySQL. Out of the 28 tables that were imported from SQL Server dump file, only one table (*TcRecords*) had all the relevant XML data in it. The column in that table was simply dumped into a text/XML file through the command line tool:

```
mysql -u avl -p -v JerwoodDiscovery < bl_xml.sql > bl.xml
```

“JerwoodDiscovery” is the database name the BL dump file was migrated into with the MySQL Workbench data migration wizard (see above), “bl_xml.sql” is the name of the file holding the SQL statement selecting the correct column in *TcRecords*: `select ItemXob from TcRecords;` and “bl.xml” is the name of the file that column data is redirected into. After issuing that command, you should have a file bl.xml which holds the relevant XML data from the table *TcRecords* from the BL SQL Server dump file. This XML needed some cleaning before being imported into BaseX, just to make the XQuery queries run smoother (and to make it easier to re-cycle the scripts created previously for other data sets):

1. Delete the first line (these are the column names)
2. Delete all `<?xml ...>` tags from every line but the first (use an editor with a global replace function); every row in *TcRecords* represents an individual XML file, hence the XML declaration in every row
3. Add “`<bl>`” as second line and “`</bl>`” as last line of the file, so that all rows are aggregated into one global entity

For this XML file a number of XQuery scripts were created to create a series of tables to translate the XML schema into a relational data model. The XQuery scripts export these into CSV files rather than SQL `insert` statements, simply because there were a lot of columns/tables that would have needed to be created manually and it was quicker to let the CSV import assign names and data types (and then clean up those if necessary).

It was noted that there were some duplicate lines in the data, but these could be easily identified within LibreOffice Calc (the equivalent to Microsoft Excel) and deleted. The CSV file was then

imported into MySQL as described above.

While *TcRecords* held the relevant data in terms of catalogue records, some of the tables in the SQL Server dump file had some necessary "auxiliary" data. The relevant tables were thus copied into the temporary schema within MySQL:

```
create table fuzzyphoto.bl_textbooks as select * from
JerwoodDiscovery.JerwoodTextbooks;

create table fuzzyphoto.bl_publishers as select * from
JerwoodDiscovery.JerwoodPublishers;

create table fuzzyphoto.bl_pibooks as select * from
JerwoodDiscovery.JerwoodPIBooks;

create table fuzzyphoto.bl_itemtypes as select * from
JerwoodDiscovery.TcItemTypes;

create table fuzzyphoto.bl_events as select * from
JerwoodDiscovery.JerwoodCVEvents;

create table fuzzyphoto.bl_departments as select * from
JerwoodDiscovery.JerwoodCVDepartments;

create table fuzzyphoto.bl_genres as select * from
JerwoodDiscovery.JerwoodCVGenres;
```

The create statement for the *TcRecords* table (as *bl_records*) is in a separate file `create_bl_records.sql`, as it was decided to slightly alter the field names to make it easier to use in follow up queries.

The cleaning was quite straight-forward, as it basically only required some "trimming" of text columns (i.e. using the `trim()` function in SQL to remove leading and trailing whitespace characters) and updating of id's between tables (in `clean_bl.sql`).

CollectionsOnline

A Java application was created to download and parse the records from the Collections Online (CO) website. However, once the records were put into the database and the NMeM data was received a comparison showed that the two were identical. There are more collections in CO, but they're usually about newer photographic collections that have little relation to the partner's data sets. Therefore, the CO data was removed from the *fuzzyphoto* db.

CultureGrid

A Java application was written to download data from the CultureGrid (CG; `http://www.culturegrid.org.uk/`) website. To query the CG website the photographer names from the ERPS and PEIB databases were used. The CG website provides the data as XML, so the single records were loaded into BaseX and transformed into SQL insert statements with XQuery queries (see above).

The data in the main table then needed some clean up, chiefly on typos etc. in names (see `clean_culturegrid.sql`). The auxiliary tables also needed attention, as most had different data stored in one field when it should have been several fields. For this, the original auxiliary table was copied (with "`_tmp`" at the end) and the separated fields inserted into it. Then the original data

was dropped and the "`*_tmp`" table was renamed to the original table's name (i.e. without "`_tmp`"). The scripts to do this all is named `upd_..._tmp.sql`

Library of Congress & Brooklyn Museum

AvL took a copy of David's downloaded data from the Library of Congress and Brooklyn Museum. The data was a bit messy in some places, as David had stored it in a MySQL schema with Latin1 encoding. However, there were names in the data (chiefly Japanese and Arabic photographers/sitters) that don't fit into Latin1, but need UTF8. AvL tried to re-convert it, but the functions within MySQL failed to do that (there seems to be no way to explicitly tell a string function in MySQL's SQL what encoding a specific column is in, other than at `CREATE` time). Thus, a list of of messed up characters was compiled and the conversion was done manually, i.e. calls to the `replace()` SQL function (see `clean_loc.sql`). Other than that, the insertion was rather straight-forward (see `insert_loc.sql`).

Metropolitan Museum NYC

The data from the Met arrived as an Excel sheet. Converting it to CSV and importing it into MySQL turned out to be without any issues. The data was very clean and needed only one update statement to set the "*alpha_sort*" column which holds the photographers names in surname, forename order (whereas the "*attribution*" column holds the name in forename surname order):

```
update fuzzyphoto.met set alpha_sort = 'N/A' where alpha_sort =
'';
```

This helps in querying the data when transferring it into LIDO. The only odd issue with the Met data is that it seems to have used the dates of the prints rather than the date of the shoot, so the event type "Print" should be used for the LIDO event table for these rows.

National Archives

The data from the National Archives was slightly messy. It turned out that the CSV file they gave us uses commas as field separator in most rows, but in some rows it uses the TAB character. Thus, importing it with the MySQL import tool as well as with LibreOffice failed at the first tries. Furthermore, in two rows misplaced TABS were found, i.e. instead of using one TAB there were three (without them being empty columns, but indeed superfluous characters). There were also misplaced 'NULL' words (i.e. the literal word, rather than the column value), which had to be deleted.

Also, extra commas were encountered. These seem to come from "empty" rows and are most likely a result of "gaps" in the National Archive's database, as much as the other issues described are a problem on their side (maybe the export function of their database client doesn't work properly).

Generally, most rows used quotation marks to indicate string values (as opposed to numerical values), but for some rows this rule was violated, i.e. there were quotation marks missing (usually only the starting quotation mark was missing). The only fix for this was to search for character strings (in a **good** text editor, in which one can search via regular expressions) that don't have a comma and a quotation mark as first characters. NB that's why the "NULL"s from above should be deleted first, otherwise they trigger that regular expression. The "uncouth" strings encountered were (just the first word of the string given here, to make it clearer; the "Original" value in the following table should be the term being searched for in your editor, the "Convert into" the replace term in the

editor):

Original	**Convert into**
,Photograph …	,”Photograph ...
“”	[null]
,Imperial …	,”Imperial ...
,Cabinet …	,”Cabinet …
,A bust …	,”A bust …
,(1) ...	,”(1) …
,Instantaneous …	,”Instantaneous …
,CDV …	,”CDV …
,Cart …	,”Cart …
, (1) …	,”(1) …
, 'Photograph …	,”'Photograph ...

The following two occurred at the end of strings:

[blank],19	“,19
[blank],18	“,18

There were some rows that were essentially empty rows. “Essentially” because although they had some value set, this was one of the following: “Item number not used”, “Number not used”, “... not registered”, “... not used” or “... Not Used”; i.e. it is not possible to link them into the National Archives collections. These rows should be (and have been) deleted.

Unfortunately, since both MySQL and LibreOffice Base/Calc won't load the data in its original state, one has to do the cleaning “manually”, i.e. in an editor. As the errors in the original CSV might be random (i.e. there might be different issues with a new export), there is little in terms of a strategy that could be offered at this point. One should try to import the data and check the messages and log files. Sometimes the MySQL importer gives only summary statements, such as “x many rows in data, y many rows imported, z many warnings”. One then needs to check whether the number of rows in the MySQL table is correct, and check the log.

The import command used was:

```
mysqlimport --fields-terminated-by=',' --lines-terminated-by='\n'
–fields-optionally-enclosed-by='”' -p -u <username> -v -L
fuzzyphoto national_archives.csv
```

Once the data was in there, the data field (there are just two usable columns, the id – which is split across two columns, see below – and the description field, which holds all information), needs to be parsed for the authors. The description field more or less only holds the photographer's name and a description of the image (sometimes the description is also the title, but seems to be rather random).

Therefore, I created an additional column for the photographer and extracted that name from the description column. Luckily, other than the NMS data, the National Archives used uniform structures and tokens in their description field, so the extraction was not overtly complicated (all in `clean_narch.sql`).

National Museums Scotland

The data of the NMS arrived in a relatively unstructured state. The Excel sheet consisted of only two columns: an id and a description field where all meta-data such as creator, date of creation, title, etc. were listed together. Moreover, the structure of this description field was not consistent. The data had been entered by volunteers, and every data entry volunteer apparently used a different structure. This made organising the data rather hard. Some data has been extracted when keywords such as “created in” were provided next to the dates and so on, but a good deal remains unstructured as part of the `lido_photography.object_description` field.

National Media Museum

We received the data from the NMeM as an Excel sheet. As a first try, it was exported into a CSV file, but the import into MySQL (in the way described above) failed. Apparently, there were newline characters in the description column that confused the importer. It was consequently decided to try it with LibreOffice's Base program, as this is usually more tolerant towards irregular characters (and can be configured to handle special characters).

To do this, do the following:

1. Create a new db in Base, use “Use existing connection” and “Spreadsheet” as type
2. Import the NMeM spreadsheet (i.e. the Excel file); NB this will only provide an editing interface to the file, and not create an actual copy of it within Base
3. Create another db in Base, use “Use existing connection” and “MySQL” as type and connect to the db “fuzzyphoto”
4. In the first db (the NMeM data) in the list of tables, do a right-click and “Copy” on the table; in the second db (the MySQL connection) go to the table list and do a right-click and “Paste Special”, pick “Data source table” as source, go through the wizard and pick the correct column data types. If you would just do the normal “Paste”, data types would be guessed (and wrongly in most cases) by the wizard and cause another error

The cleaning up process was more or less solely about the names of the creators. In the beginning there was only one column for creators and multiple creators were concatenated into that one column. A quick check revealed that there were some entries with more than four names in that one column, so three more columns for extra names were added to the table. Then, a number of update statements were run to split the one original creator name column into these added columns. Furthermore, some updates were run to get some consistency into the spelling of the names (see `update_maker.sql`).

Victoria & Albert Museum

The data from the V&A was received as XML data and was well structured. The format was discussed with the V&A prior to the delivery, upon a sample set sent earlier. There were five objects in the XML schema that needed to be separated out into individual tables in the MySQL db (see the `create_*.sql` files). There were some empty nodes in the XML file, but they were quickly

spotted and deleted in XQuery, so that they posed no problem when creating the SQL insert statements from within BaseX.

Cleaning up the data involved straightening out some inconsistencies in the creator names, such as replacing HTML entities (e.g. "`&`") or checking for multiple creators in one data field. Furthermore, fields were set to null when they held a string of length zero in various tables.

Sorting out multiple titles in the one title field was somewhat more elaborate. Some items had multiple titles, often the photo and the album title. In these cases it made more sense to use the brief description field as title. To achieve this all titles from the *vanda_titles* table that are multiple titles to one id were dropped, then the unique titles were inserted into the main table, and then the main table was updated with the brief descriptions for those rows that had been dropped (i.e. the ones with multiple titles).

There are multiple dimensions for the image and the mounted image. Under the LIDO schema, this would result in different events, i.e. several rows. However, there were no dates etc. given for the mount, so insufficient data was available to create an extra event. (Although events in LIDO do not need to have a date, the fuzzy logic algorithms would need more than just different dimensions). The mount dimensions were therefore dropped from the data set.

St Andrews University Library

The data from St Andrews came as a bunch of CSV files (13 CSV files, and one `ini` file with meta-data on the CSV file, and an Excel sheet with a diagram showing the relations between the different entities). The tables were created with the help of the `ini` file, i.e. the SQL `create` statements were informed by the `ini` file (see `create_standrews_tables.sql`). The CSV files were then imported by the method described in the General Issues sections.

The data was pretty clean (St Andrews spent quite some time to clean it up themselves), so it was essentially only a couple of typos that needed corrections.

Musee d'Orsay

The MO sample data was delivered as an XML file. A number of XQuery queries turned this into SQL insert statements, with the table create statements being authored manually (see `create_*.sql` files). The only peculiarity with the data was that there were several entities with "*historique*" in their names. The difference in name was the only variation, the structure being the same across all these entities. Consequently, only one table in SQL was created as *orsay_historique* and the different entities were imported into this one table, with a column "*hist_type*" that would specify from which of the different entities the rows came. Similarly, there were two entities with "*actuelle*" in their names, for which one table *orsay_situation_actuelle* was created in MySQL.

The data as such was quite clean. The only update necessary was deleting "empty" rows, i.e. rows which for some reasons had no actual data in it (NB this became quite an issue with some of the tables in the final dataset we got from them, making some of the SQL files quite large (~60MB) with a lot of empty rows that get imported, just to be dropped afterwards, wasting time; as an improvement, the XQuery script should filter out those rows immediately before the SQL export; on the other hand, not all of the tables/files were actually used for the LIDO schema – see below – so some of the SQL files can be ignored, thus saving time).

Transferring the partners' data into the LIDO schema

Entities and relationships

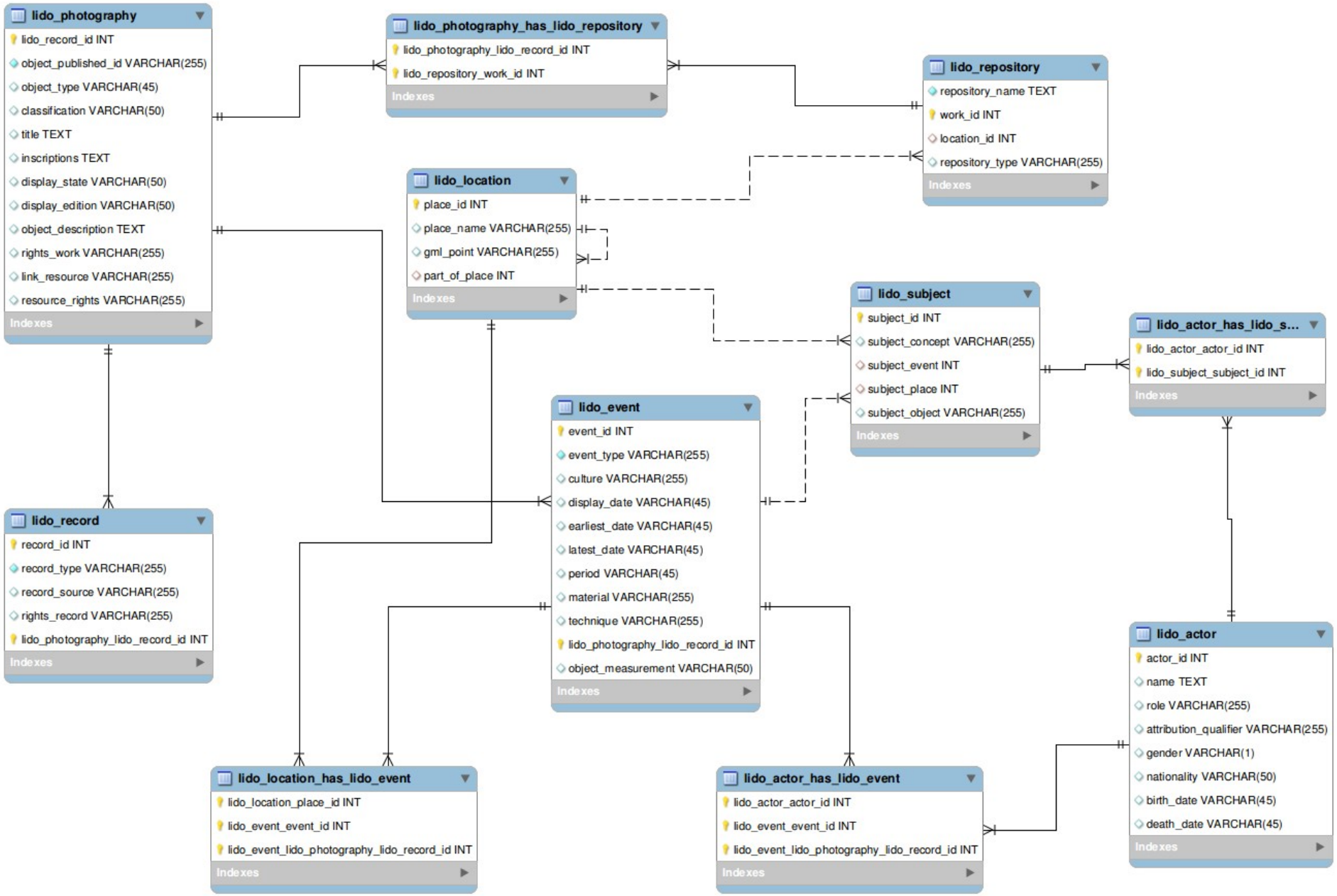
LIDO is defined as an XML Schema Definition (XSD; see `http://www.lido-schema.org/schema/v1.0/lido-v1.0.xsd`) and is thus not immediately usable with MySQL (unless one wants to use the XML features of MySQL). The XML schema was therefore manually translated into an Entity-Relationship-Model (ERM), based on the schema visualizations on the LIDO webpage. The only deviation from the XSD is that the field "*Object Measurement*" was put in the event entity rather than the main record entity. This is because of different sizes of photographic prints made from the same photograph, with prints being regarded as events to one original photo and not new objects (as a LIDO introductory paper itself suggests). Consequently, measurements change with every event, a.k.a. print, and the events table needs a column for that. Furthermore, a column named "*internal_remark*" was added to the main table (*lido_photography*) to mark the origin of the data (i.e. strings like "nmem" etc. were used to differentiate between the sources of data). While the different partners could be distinguished by using the *lido_repository* table, this would involve several Inner Joins. For the cleaning operations, using *internal_remark* was much quicker.

The SQL scripts to create and clean the LIDO schema in MySQL are all in the lido/ subdirectory (see below for a ERM diagram of the LIDO schema implementation). The SQL scripts to migrate the individual data sets from the partners reside in the respective partner's subdirectory (as `insert_lido_*.sql`). The analysis of what columns to transfer across between an individual data set was entirely manual, i.e. the respective data was visually inspected and upon that a decision was made about the corresponding columns in the partner's data and LIDO. When transferring creator names into the table/field *lido_actor.name*, the field *lido_actor.role* should usually be "Photographer", unless the data clearly states/shows that it is a different role (such as "Printer"). The nomenclature for the names should be <*surname*>, <*first name*> <*middle name*>.[4] Some LIDO data has to be "made up" as it is not in the data from the partners as such, e.g. the data for the *lido_repository* table, which is data about the respective partner institution etc (see the SQL scripts).

To implement the LIDO schema, the XSD diagrams were used as a blue-print and the ERM was modelled in the modeller in MySQL Workbench. This made it possible to draw an ERM graphically, define data types and relationships, and then have the software generate the database in SQL. The default name for that database assigned by the modeller is "*mydb*" and it was hard to figure out how to change that, so it was left at that.

The ERM diagram for the LIDO schema used for the FuzzyPhoto project is as follows:

[4]It should go without saying that the transfer into the LIDO schema for the names should be done with `INSERT IGNORE INTO … SELECT DISTINCT …` to avoid duplicates.


lido_photography
lido_record_id INT
object_published_id VARCHAR(255)
object_type VARCHAR(45)
classification VARCHAR(50)
title TEXT
inscriptions TEXT
display_state VARCHAR(50)
display_edition VARCHAR(50)
object_description TEXT
rights_work VARCHAR(255)
link_resource VARCHAR(255)
resource_rights VARCHAR(255)
Indexes
lido_photography_has_lido_repository
lido_photography_lido_record_id INT
lido_repository_work_id INT
Indexes
lido_repository
repository_name TEXT
work_id INT
location_id INT
repository_type VARCHAR(255)
Indexes
lido_location
place_id INT
place_name VARCHAR(255)
gml_point VARCHAR(255)
part_of_place INT
Indexes
lido_subject
subject_id INT
subject_concept VARCHAR(255)
subject_event INT
subject_place INT
subject_object VARCHAR(255)
Indexes
lido_actor_has_lido_s...
lido_actor_actor_id INT
lido_subject_subject_id INT
Indexes
lido_event
event_id INT
event_type VARCHAR(255)
culture VARCHAR(255)
display_date VARCHAR(45)
earliest_date VARCHAR(45)
latest_date VARCHAR(45)
period VARCHAR(45)
material VARCHAR(255)
technique VARCHAR(255)
lido_photography_lido_record_id INT
object_measurement VARCHAR(50)
Indexes
lido_record
record_id INT
record_type VARCHAR(255)
record_source VARCHAR(255)
rights_record VARCHAR(255)
lido_photography_lido_record_id INT
Indexes
lido_actor
actor_id INT
name TEXT
role VARCHAR(255)
attribution_qualifier VARCHAR(255)
gender VARCHAR(1)
nationality VARCHAR(50)
birth_date VARCHAR(45)
death_date VARCHAR(45)
Indexes
lido_location_has_lido_event
lido_location_place_id INT
lido_event_event_id INT
lido_event_lido_photography_lido_record_id INT
Indexes
lido_actor_has_lido_event
lido_actor_actor_id INT
lido_event_event_id INT
lido_event_lido_photography_lido_record_id INT
Indexes

The transfer schemas were as follows:

Birmingham City Library

Only one temporary table for BCL, which needs to be split up to fill various LIDO tables.

Temporary Table	LIDO
date_range	lido_event.earliest_date/latest_date (split)
Format	lido_event.material
Dimensions	lido_event.object_measurement
Title	lido_photography.title
Description	lido_photography.object_description
Reference	lido_photography.object_published_id
Creator	lido_actor.name
	lido_event.event_type = 'Shoot'

British Library

The data from the BL is spread across 15 tables, but a lot of this isn't used (on average its one column per table). All tables are linked through the “id” column.

Temporary Tables	LIDO
bl_records.RecordId	lido_photography.object_published_id
bl_records.Caption	lido_photography.title
bl_records.StartDateRange	lido_event.earliest_date
bl_records.EndDateRange	lido_event.latest_date
bl_records.Height/Width	lido_event.object_measurement
bl_records.Process	lido_event.technique
bl_records.DescriptiveNotes	lido_photography.object_description
bl_records.CopyrightStatus	lido_photography.resource_rights
bl_records.Subject	lido_subject.subject_object (not used, since only one other data sets has this information)
bl_records.Location	lido_location.place_name (not transferred yet, as it requires more work)
bl_records.Inscription	lido_photography.inscription
bl_secondarysupport.support	lido_event.material
bl_events.Event	lido_subject.subject_object (see above)
bl_genres.Genrer	lido_subject.subject_object (see above)
bl_photographers.name	lido_actor.name (role = 'Photographer')
bl_portraitsubject.name	lido_actor.name (role = 'Portrait Subject')
	lido_event.event_type = 'Shoot'

CultureGrid

The data downloaded from the CultureGrid web service is spread across eight tables, although little data from each table is actually used (all tables are linked through the “aggregator_internal_id” column).

Temporary Tables	LIDO
culturegrid.authority_name	lido_photography.resource_rights
culturegrid.dc_description	lido_photography.object_description
culturegrid.dc_identifier	lido_photography.link_resource
culturegrid.dc_rights	lido_photography.rights_work
culturegrid.dc_title	lido_photography.title
culturegrid.temporal_from	lido_event.earliest_date
culturegrid.temporal_to	lido_event.latest_date
dc_creator.dc_creator	lido_actor.name
dc_subject.dc_subject	lido_subject_object (not transferred, since no other data set has this kind of information)
dcterms_spatial.dcterms_spatial	lido_location.place_name (not transferred yet, as it requires some extra work to get it in the right format)

ERPS & PEIB

The Photographic Exhibition In Britain (PEIB) and the Exhibitions of the Royal Photographic Society (ERPS) databases, created at DMU, were transferred into the LIDO schema. The transfer was quite straight-forward; however, given that these databases have digitized catalogues from exhibitions in them, rather than actual records of actual collections of historic photos, the structure is slightly different. For the LIDO schema this simply means a different event type and less data (as there is no repository for these datasets, for example).

The only real issue were the exhibitor names; names were given in full with life dates in brackets as part of the name field. The dates had to be stripped to make it LIDO compliant.

PEIB	LIDO
exhibits.exhibid	lido_photography.object_published_id
exhibits.exhibtitlenorm	lido_photography.title
concat(exhbndetail.opened, ' ', exhbndetail.year)	lido_event.earliest_date
concat(exhbndetail.closed, ' ', exhbndetail.year)	lido_event.latest_date
exhibits.processnorm	lido_event.technique
exhibits.photographernorm	lido_actor.name
	lido_photography.event_type = 'exhibition'

ERPS	LIDO
exhibitors.name	lido_actor.name

exhibits.etid	lido_photography.object_published_id
exhibits.title	lido_photography.title
exhibitions.datestart	lido_event.earliest_date
exhibitions.dateend	lido_event.latest_date
exhibits.process	lido_event.technique
	lido_event.event_type = 'exhibition'

Library of Congress and Brooklyn Museum

Data was across tables, but David created a view to unify them, which is what was copied across.

Temporary Table	LIDO
uid	lido_photography.object_published_id
title	lido_photography.title
description	lido_photography.object_description
person	lido_actor.name
process	lido_event.technique
date	lido_event.earliest_date/latest_date
source	lido_photography.internal_remark
image	lido_photography.resource
	lido_event.event_type = 'Shoot'

National Archives

The data from the National Archives UK was in one table. After the CSV file was inserted one column name had to be altered, because it had a slash character in it, which might have proven troublesome for queries.

Temporary Table	LIDO
series + '_' + piece_ref + '_' + item_ref	lido_photography.object_published_id
Content [renamed from scope/content]	lido_photography.object_description
date_text	lido_event.display_date
	lido_event.event_type = 'Shoot'

National Media Museum

Only one temporary table for NMeM.

Temporary Table	LIDO
id_number	lido_photography.object_published_id
item_name	lido_event.technique
title	lido_photography.title
maker	lido_actor.name

date_made	lido_event.earliest_date (only one date)
measurements	lido_event.object_measurement
description	lido_photography.object_description
	lido_event.event_type = 'Shoot'

National Museum of Scotland

As mentioned, the NMS data is relatively unstructured, but some data could be extracted. The fields are as follows:

Temporary Table	LIDO
id	lido_photography.object_published_id
description	lido_photography.object_description
type	lido_event.technique
maker	lido_actor.name (role = 'Photographer)
title	lido_photography.title
dimensions	lido_event.object_measurement
	lido_event.event_type = 'Shoot'

Metropolitan Museum

The MET data is one temporary table only. The field "*alpha_sort*" is used for the creators' name rather than the field "*attribution*", as the former is in the right format (i.e. surname first) and identical to the latter otherwise.

Temporary Table	LIDO
object_id	lido_photography.object_published_id
alpha_sort	lido_actor.name
object_name	lido_photography.object_type
Title	lido_photography.title
begin_date	lido_event.earliest_date
Medium	lido_event.technique
Description	lido_photography.object_description
	lido_event.event_type = 'Print'
	lido_photography.link_resource = 'http://www.metmuseum.org/Collections/search-the-collections/1900' + object_id

Museé d'Orsay

The data from the Musee d'Orsay is held in 14 tables, although it turned out that at least three of them can be ignored, as the data in them is not usable for the project (or no other partner has delivered this kind of data, so it is slightly pointless to include it). Since loading of the SQL files can be quite time-consuming, one should probably not even import the SQL files, speeding up the

process. The tables not being used were *orsay_iconographies* (subject terms), *orsay_historique* (items' histories, e.g. date of acquisition etc.) and *orsay_fonds* (sponsors). The SQL scripts all start with `insert_*<table name>.sql`.

The Musee d'Orsay data was very clean and their data model quite sophisticated. There are three areas were care must be taken: 1) The data has often more than one title (up to three) per item, usually a short and a long version, but also sometimes the French and the English title. All of the title variants might be relevant, yet LIDO only allows one title (or at least in our SQL implementation of it). As a solution, and considering that this is the only dataset having this issue, the different titles were concatenated into one line, with the different titles separated by a vertical bar (“|”) if applicable. 2) A similar issue arose with the inventory number of the item, where there are up to four in the Orsay data. Maybe this indicates different prints of the same image, which in LIDO should be separated out in four different events (of type 'Print'), but there is no other data to go along with it, such as the date of the individual print. So, again, the different inventory numbers were concatenated into one string, and the different numbers separated by an ampersand (“&”). (NB the opposite, unfortunately, is also true for the data: there are records that have several internal id numbers for the same inventory number, and the only obvious difference are different inscriptions, i.e. title and dimensions are the same. During a team meeting it was decided to delete these duplicates, and only enter one record with one inventory number). 3) For the dates, the Musee d'Orsay introduced different qualifiers to denote whether it is a certified date, a date interval or an open interval (i.e. it is given a start or end date to say whether a photo is believed to have been taken before/after a certain date). To hold this in LIDO, the following conventions were used: a) if both dates (*earliest_date/latest_date* in *lido_event*) are given, these are exact intervals, i.e. the shooting was between these dates (only years are given, so there are a good number of rows with the same start and end date/year, specifying that the shooting was in that year). b) if either the start or end date is NULL, it is an open interval.

It should also be noted, that there are no image descriptions in the Orsay data, but a column with inscriptions that were made on the photo. In other datasets inscriptions are often recorded with the image description. On the other hand, there is a column for inscriptions in LIDO, as well as for descriptions. For the Orsay data, the insciption column is used, while no other dataset has anything in there.

Furthermore, while there are dimensions in the *orsay* table, it is in a different format, so they had to be converted (see `clean_orsay.sql`). As another issue, the Musee d'Orsay seems to have photos in their store that they don't own, i.e. they are conserving/safekeeping photos from other museums. Consequently, for every item in *lido_photography* from the Musee d'Orsay data there are two repositories for it: the 'Current' one (i.e. where the photo is held at the moment) and the 'Owner' (i.e. the legal owner of the photo, regardless where the photo is stored). This may be relevant for the widget, as interested persons may need to contact both repositories to get access to a certain image.

The tables are linked via the “*id*” column, and “*num_fiche*” in the main table (*orsay*), respectively.

Temporary Tables	LIDO
orsay.copyright_id	lido_photography.resource_right
orsay.dimensions	lido_event.dimensions
orsay.designation_redigee	lido_event.technique
orsay_titres.titre (see above for comment)	lido_photography.title
orsay_situation_actuelle.mtext	lido_repository.repository_name
orsay_inventaires.inventair (see above for comment)	lido_photography.object_published_id

orsay_inscriptions.inscription	lido_photography.inscriptions
orsay_date_modele.date_debut	lido_event.earliest_date
orsay_date_modele.date_fin	lido_event.latest_date
orsay_auteurs_tmp.artiste	lido_actor.name

St Andrews University Library

The data from St Andrews is spread across 14 tables, of which only few rows were actually transferred. The tables are all linked by the "*ecatalogue_key*" column.

Temporary Tables	LIDO
st_andrews_col_artis.NamCitedName	lido_actor.name (role = 'Photographer')
st_andrews_col_artis.BioBirthDate	lido_actor.birth_date
st_andrews_col_artis.BioDeathDate	lido_actor.death_date
st_andrews_ecatalog.ColObjectNumber	lido_photography.object_published_id
st_andrews_ecatalog.PhoRecordLevel	lido_photography.object_type
st_andrews_ecatalog.PhoOriginalDateEarliest	lido_event_earliest_date
st_andrews_ecatalog.PhoOriginalDateLatest	lido_event.latest_date
st_andrews_ecatalog.ColMainTitle	lido_photography.title
st_andrews_ecatalog.ImaDescription	lido_photography.object_description
st_andrews_pho_funct.PhoFunctionType	lido_subject.subject_object (not transferred, as only one other data set has this data)
st_andrews_pho_media.PhoMedia	lido_event.technique
st_andrews_sub_subject.SubSubjects	lido_subject.subject_object (see above)
st_andrews_sit_site_r.SummaryDate	lido_location.place_name
st_andrews_pho_sitte.NamCitedName	lido_actor.name (role = 'Portrait Subject')
st_andrews_pho_sitte.BioBirthDate	lido_actor.birth_date
st_andrews_pho_sitte.BioDeathDate	lido_actor.death_date
	lido_event.event_type = 'Shoot'

Victoria and Albert Museum

The data from the V&A is distributed across five tables (all linked through the "*id*" column in each).

Temporary Tables	LIDO
vanda.museum_no	lido_photography.object_published_id
vanda.spec_phys_desc (if empty use spec_brief_desc)	lido_photography.object_description
vanda.mus_mat_note	lido_event.material
vanda.url	lido_photography.link_resource
vanda.spec_obj_prod_date_start	lido_event.earliest_date

vanda.spec_obj_prod_date_end	lido_event.latest_date
vanda_dimensions.name/width/height	lido_event.object_measurement (concatenate width and height, use only name = “images”)
vanda_obj_names.title	lido_event.technique
vanda_titles.title	lido_photography.title

Cleaning the data in LIDO

There are a number of SQL and Python scripts to clean/unify the meta-data once all the partners' data has been migrated into the LIDO schema. There are also some checking queries and some general queries in the lido/ subdirectory, which should be self-explanatory.

Duplicates in the actor table were detected and cleaned up in SQL. However, this was only done for straight matches (obvious typos were cleaned before), i.e. spelling variants were ignored. This was chiefly done because it would have involved quite some extra work at this point, and the fuzzy logic algorithms should be able to pick this up at a later stage anyway. Furthermore, the actor table has a column for gender in it. Given that sometimes the only information about the creator of a photo in the partners' meta-data was a society or company, the following codes were used: *M* for male, *F* for female, *C* for company (or society), and *U* for unknown. Sometimes only a surname was present in the meta-data, so that no gender could be determined, hence the code *U*. When several names were used in the creator field in the partners' data (e.g. “Smith & Wesson”) this was regarded as code *C*, although judicially it may not have been a company but just two photographers cooperating. Since most creator's in the partner's meta-data didn't have any information on their gender, the i-gender Web API (http://www.i-gender.com/main/index.html) was used to update the actor table in LIDO (with `upd_gender.py`). Some had to be manually checked (such as the unknown ones or the companies). Sometimes the i-gender engine would fail and produce nonsense (i.e. identifying a name as female when it is clearly male, and vice-versa), these had to be manually altered. i.e. proof-reading the result of that Python script is inevitable.

A big issue were different formats for recording dates and duration. In total, twenty different date/duration formats have been encountered across the partner's data. Care has been taken to convert them into a unified date format (ISO8601) using regular expressions, but some formats escaped such conversion. These were chiefly 'fuzzy' dates, such as “early 19th century”. To handle this in an efficient manner, the following strategy was chosen: if there were proper dates, the columns *lido_event.earliest_date* and *lido_event.latest_date* were used (sometimes only one of the two could be set, as there was only present in the source data). If there was a 'fuzzy' date, this values was copied into the column *lido_event.display_date*. This might not be in the LIDO spirit (i.e. this is not what that column was intended to hold), but the only alternative would have been to add a column to hold 'fuzzy' dates, which would stray from the LIDO standard even more.

The following issues were detected after all the partners' data was transferred into the LIDO schema:

1. 10,000 records in *lido_photography* had no corresponding entry in *lido_event*, which should never happen (there always needs to be at least one event such as “Shoot” or “Print”, even when dates etc are not given). Analysis: All these records were from the St Andrews data set; the problem is that the query to transfer the St Andrews data to LIDO inner-joins data from two temporary tables and the number of rows in both tables are not identical. Solution: use Left Join when doing the query. Now fixed.
2. 1,187 records in *lido_photography* have a duplicate “*object_published_id*”. Analysis 1: 308 rows are in the CultureGrid data that are also in the NMeM data; Solution: delete the

CultureGrid ones. Analysis 2: the other rows seem to be multiple photos in an album all being given the same object id (from St Andrews and V&A data); Solution: this is ok, ignore.

3. 58 rows with `object_published_id = ''` (i.e. not such id). Analysis: they are all from the BCL data set, checked the temporary table and original data file: most of them have no description and most not even a title. Solution: delete.
4. 2,989 rows had neither title nor description and are thus useless. Solution: delete.
5. While care has been taken not to duplicate photographers, it is inevitable that some are duplicated simply through spelling variants, i.e. typos in the partners' data. A lot could be sorted out with SQL and Python scripts, but there are certainly a good number of them present. The fuzzy logic algorithms should be able to cope with those, i.e. to identify them as matches regardless.

General Issues

Data transfer

While trying to copy the LIDO database from AvL's local machine (MySQL 5.5) to KMD2 (MySQL 5.0.26) several issues arose:

1. The dump file from MySQL 5.5, although plain SQL statements, turned out to be incompatible with MySQL 5.0; there is a switch for `mysqldump` to force downward compatibility.
2. The dump file needs to be transferred to KMD2 and imported locally, as AvL's machine and KMD2 are on different subnets and DMU's routers seem to block port 3306 (which MySQL runs on).
3. There is a bug in MySQL before version 5.0.48 that crashes the database (not only when importing data), to do with a flaw in index truncation causing an assertion in MySQL's code to fail (see `http://bugs.mysql.com/bug.php?id=28125`). Ideally, the MySQL version on KMD2/3 should be updated, but the SLES10SP2 version of Linux running on these machines don't seem to have a newer package available in their repository (most likely because SuSE has discontinued support for this version of their Linux distribution). Packages might be available on OpenSuSE or older mirrors, such as `http://ftp5.gwdg.de/pub/linux/suse/opensuse/discontinued/distribution/10.2/repo/`. However, given time predicaments etc., AvL decided not to update the MySQL on KMD2.

The commands to successfully migrate the LIDO database from MySQL 5.5 to MySQL 5.0.26 on KMD2 were thus ('mydb' being the schema the LIDO database is held in):

1. First, the index on the column *lido_photography.title* was dropped entirely, as it is likely to be the trigger for above mentioned bug. It can be re-created after the import, or it can be ignored. Its main use was when the data was cleaned up, to help with queries.
2. On the machine with MySQL 5.5: `mysqldump -u <db_user_name> -p --add-drop-database --add-drop-table --compatible=mysql40 --complete-insert --set-charset --disable-keys mydb > dump.sql`

3. (S)FTP the file `dump.sql` to KMD2 (or whatever machine you are using as database server).
4. On the designated server, import the file `dump.sql` like this: `mysql -u <db_user_name> -p < dump.sql`

NB you may have to edit the dump file and add the following lines to the top of the file:

```
drop database if exists mydb;
create database mydb character set 'utf8';
use mydb;
```

This should have actually been done by `mysqldump`, but maybe the downward compatibility switch suppressed this.

Migrating Temporary Tables to the LIDO schema

Given that some data sets arrived at about the same time, one could contemplate to copy data from the respective temporary tables into the LIDO schema in parallel. Unfortunately, MySQL capability to run queries in parallel is somewhat limited. Even for SELECT statements MySQL locks the whole table for write-access. Since LIDO uses a number of n:m relationships data is pulled in from other tables from the LIDO schema when populating those relationship tables, meaning that inserting into any of the affected tables can't have new data inserted into them while there is a SELECT query running against them (see `http://thushw.blogspot.co.uk/2010/11/mysql-deadlocks-with-concurrent-inserts.html` for an introduction into the issue). As some of those queries to build the n:m relationships can be quite slow, building the LIDO schema can take several days. This should be considered for future builds. NB make sure there are proper indexes in place in both the LIDO tables and the temporary tables to speed up the migration.

Transferring data to the cluster server

One issue arose when trying to copy the LIDO data (applies to the temporary tables as well) to the cluster server.[5] The default database engine type for MySQL databases is “InnoDB”, whereas for clustered servers, this has to be “NDB” (check the MySQL documentation to learn about the differences). However, the NDB database type has a row size limit of 14,000 bytes per column. The only column large enough to trigger this is the *description* field (in *lido_photography.object_description*, or in the respective temporary tables). Since we’re using UTF-8 encoding, which uses 3 bytes per character (i.e. there’s a 4,666 character limit, c. an A4 page), some of the description fields would go beyond this limit. In the present version there 12 rows in the CultureGrid data that would violate the 14,000 byte restriction. Upon inspection of the rows in question were simply chopped down to comply with this limit (the descriptions in those rows looked like full-blown articles, with “Conclusions” and “References” sections in them):

```
Update fuzzyphoto.culturegrid set dc_description =
substr(dc_description, 1, 13999) where length(dc_description) >
14000;
```

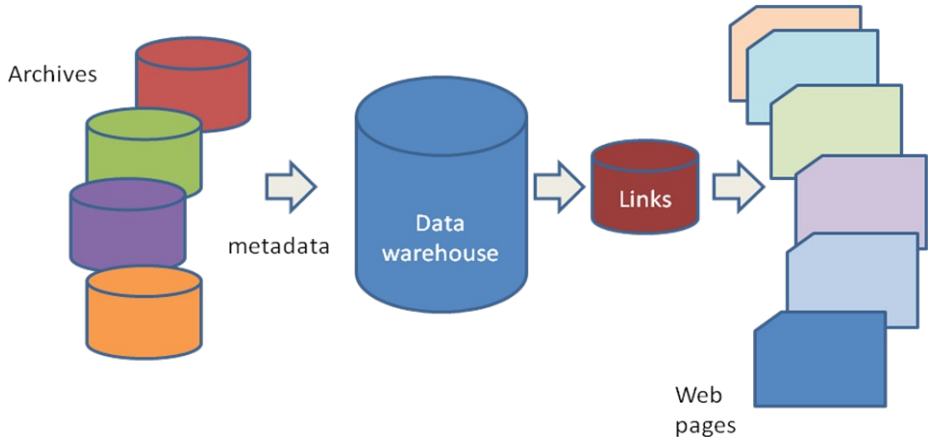
[5] While the issue arose through the CultureGrid data, future data sets might trigger it as well, so it is discussed here, rather than in the CultureGrid section.

NB This needs to be done for *lido_photography.object_description* as well, if the data has already gone in. Furthermore, the parameters for uploading the data to the cluster need to be adhusted, but this is detailed in the Cluster work package report.

Appendix 8. Specimen memorandum of understanding.

FuzzyPhoto
Memorandum of Understanding

FuzzyPhoto is a two year AHRC funded research project (AH/J004367/1) that is developing and testing computer-based "finding aids" that can recommend potential matches between Incomplete historical data sets containing imprecise information, even where there is not a precise match. It is centred around a pair of databases that together comprise the single most comprehensive record of British photographic exhibitions in the nineteenth and early twentieth centuries: Photographs Exhibited in Britain (http://peib.dmu.ac.uk) and Exhibitions of the Royal Photographic Society (http://erps.dmu.ac.uk). Early exhibition catalogues were often devoid of pictures, relying instead on written descriptions of the image or artists' impressions. This project is developing and testing computer based methods for extracting image object metadata from partners' collections, storing it in a data warehouse where it can be searched offline for potential matches. The results are stored in a second "links" data base which is interrogatable by a "widget" embedded in partners Web sites.



When visitors to any one of these sites drill down to the level of an individual image object, the widget will offer them a selection of hyperlinks to potentially related records held in other partners' collections. Visitors can choose which links, if any, they wish to follow back to the originating partners' collections.

I, (print name)……
On behalf of (print organisation name) ……
hereby give consent to the use of our catalogue records supplied to De Montfort University for the AHRC *FuzzyPhoto* project (AH/J004367/1), and hereby give all consents necessary for the extraction of metadata from these records, generation and publication of links between these records and those contributed by other organisations in the project, without time limit throughout the universe by all means and media (whether now known or hereafter invented) without liability.

I, PROFESSOR STEPHEN BROWN, on behalf of DE MONTFORT UNIVERSITY, hereby give an undertaking that the AHRC *FuzzyPhoto* project (AH/J004367/1) will not republish any of the partners' images or their metadata, will keep all catalogue data supplied by project partners confidential, will not use this data for any purposes other than the *FuzzyPhoto* project and will not seek to exploit this data commercially.

Signed	Signed	
Organisation	Organisation	De Montfort University
Date	Date	

Appendix 9. WP 3 Batch Loader report

FuzzyPhoto AHRC AH/J004367/1

Work Package 3 Report: Batch Loader

Dr. Jethro Shell, Mr. David Croft

25th October, 2013

Status: Draft
Distribution Type: Public

Keywords: Batch Loader, MySQL, Python, Database, Web Service

Contents

1. Introduction

It was identified within the project structure that alongside the incorporation of data from the partner organisations directly donated, to maintain sustainability there would be a need for a consistent system of data upload on a periodic basis. A review was carried out (See report Batch Loader Requirements Report v1.1) to summarise the basic needs of such a system. This report discusses the development of a batch loader to import additional data from three key partner organisations from within the interest group.

Outputs of this work package comprimises:

- This report.
- A Python based implementation for the automated extraction and uploading of data to the FuzzyPhoto database cluster.

The elapsed time for this work package was two months. The resources required to complete this work package were 24 person-days.

2. Outlined Batch Loader Structure

The prime goal of the project is to produce links between images maintained across partner websites and additionally sourced data. To sustain the relevance of this data, there was identified a need to update this information on a periodic basis. The raw data supplied produces issues regarding consistency and clarity. This is discussed in depth within the report for Work Package 3. Based upon the findings in Work Package 3, a decision was taken to pursue data sources that were persistently more consistent throughout. The chosen data is to come from sources at the Victoria and Albert Museum, University of St. Andrews, and The Library of Birmingham.

The outline batch loader structure consists of:

- Uploading of data.
- Identification of correct and relevant source.
- Transfer to defined database format.
- Upload to LIDO database.

This outline structure is shown in Figure 1. The overall structure of the batch loader follows that outlined in Batch Loader Requirements Report v1.1. In this report two approaches were proposed: an individualised approach and a generalised, modular approach. Following the analysis of the data, it was decided that a more robust process would result from the use of a individualised configuration. The individualised approach constructs different processes for each institution in order to upload, identify, transfer and convert the data to the final database. The speed of implementation is proportional to the quantity of institutions incorporated. As this is only three, this method can supply both a robust methodology with minimal development time. The static nature of the method does not, however, allow for changes to the partner institutions data structures. This needs to be taken into consideration. The following sections take each part of the batch loader structure in turn, providing a discussion of the process.

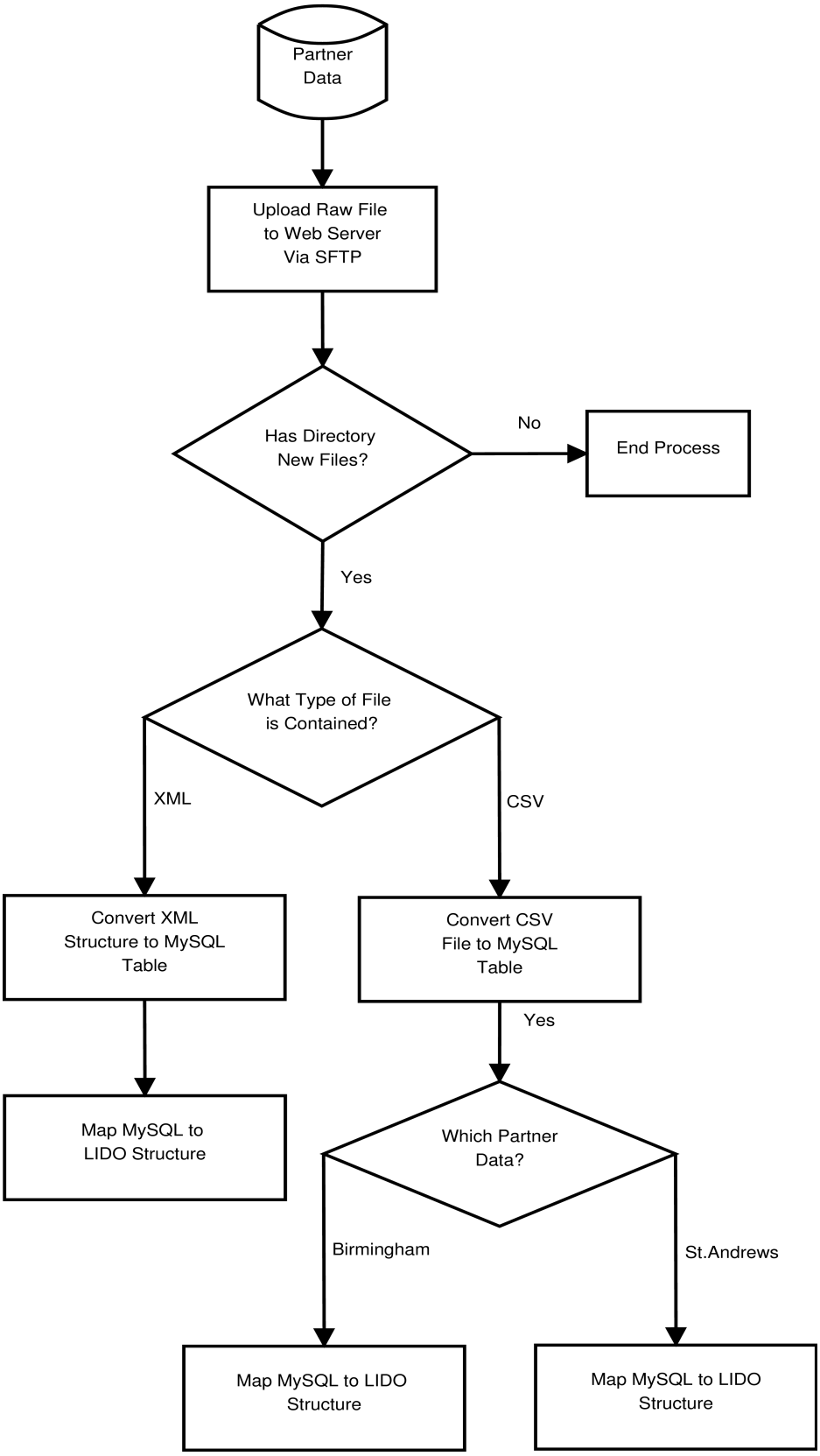


Figure 1. Outline of the Batch Loader Process.

2.1 Upload

In order for new data to be accessible to the batch loader for processing, it initially is loaded to a server through the use of Secure File Transfer Protocol (SFTP). SFTP allows the

transferring of files from a client (partner) to a host (Fuzzyphoto project server) over a connection with increased security. The SFTP is configured to use Public-key cryptography. Public-key cryptography uses two separate keys, one which is private and one which is public. The public key is used to encrypt plaintext or to verify a digital signature with the the private key being used to decrypt ciphertext or create a digital signature. This method needs to be established on both the client and host units.

SFTP also allows the files to be transferred with basic attributes such as timestamps, an important aspect of this process and an advantage over common File Transfer Protocol (FTP) which does not provision for this.

Each institution will add data on a periodic basis (currently standing at 6 months) to the server. The data is subsequently transferred to the data cluster to be processed.

2.2 Identification

The construction of the batch loader application used the Python programming language. A flexible, adaptable language, Python contains numerous libraries that allow for the processing of language and incorporation of SQL statements, both of which are integral to the batch loader. A brief discussion of how to run the application is given in the Appendix of this document.

The first stage of the main application is to identify the contents of the directories uploaded. The batch loader will periodically run to check the contents of the directories. This process looks to identify a number of key elements:

8. Is there data in the directories?
9. Is the file of the correct type for that institution?
10. Is the file uploaded younger than previous data uploaded?

2.2.1 Data Contained in Directories

To identify whether there are any contents within the directories, each subdirectory of the defined directory is scanned. Any files that are contained within the directories are recorded. If the directories are empty, the whole process is bypassed. This is accomplished through the use of the Python `os.path` and `directory` methods which allow for the navigation through directories. In the batch loader, the method `_Check__dirName` in class `upload` is used.

2.2.2 Correct File Type

If the directories are found to contain data, each file is processed to ascertain its type. Each institution has a predefined file type associated with its process I.E XML is associated with the Victoria and Albert Museum. This allows for a robust process to be employed. To gain the types, each file is examined using its extension. If the extension does not relate to the specified type, the file is not used. This processes uses the Python `os.path.` In the batch loader, the method `_Check__fileType` in class `upload` is used.

2.2.3 Compare Timestamps of Files

The comparison of the files within the directories has dual importance. Files may remain within the directory with the same name from previous uploads, also partner institutions may upload old files. Both cases need to be taken into account.

To accomplish this, firstly the timestamp of the uploaded file is gained using the `_Check__filename`, `_Check_getTimeStamp` and `_Check_timeStampUpload`

methods from the `upload` class. The filename is gained from the file without the inclusion of its extension. This is passed to the `_Check_timeStampUpload` method along with the timestamp, acquired using the `_Check_getTimeStamp` method. To compare previous files, a CSV file of timestamps mapped to the appropriate file name is contained within the configuration. The `_Check_timeStampUpload` compares the information supplied to data within the file. Based on this data, the method produces a `true` value if the uploaded file is younger.

Based on the output from the method, the `_Check__uploadTimes` confirms the file, and adds the new timestamp to the configuration file removing the previous timestamp. If the file is older than the data contained in the configuration file, the subsequent processes are skipped.

2.2 Upload

Each of the raw sources provided by the partner organisations offer unique problems in terms of processing and cleaning of data. For an in depth discussion of the problems encountered, and solutions provided to this process, see the Work Package Report 3. Due to the nature of the raw data provided, three of the partner organisations have been focussed on. The batch loader approached the uploading of each of the datasets in differing manners due to the nature of the raw data.

2.2.1 Library of Birmingham

The Library of Birmingham data is provided in a raw form as an excel sheet. To provide a consistent format, it will be requested that they provide a Comma-separated Values (CSV) file. The Birmingham data is formed from a single table that consists of eight columns. In line with previous work carried out in Work Package 3, a single MySQL table is constructed.

To be able to revert any changes made to the data and provide a MySQL structure to work from, the uploading process loads the raw data to MySQL interim tables. As the Birmingham data is a CSV file (currently Excel), the process to upload this format can use methods within the standard MySQL structure.

The partners chosen to provide continued data support were specifically sort due to the cleanliness of the source. Whilst the Birmingham data provided requires minimal cleaning, the data format was shown to require pre-processing before entering the LIDO format. This will be discussed further in later sections.

Each of the institutions data is loaded into a specific directory. As was previously discussed, this directory is processed to check for the correct file type. The Library of Birmingham data is to be supplied as a CSV file. The final structure of the Fuzzyphoto database is an entity relationship MySQL form. To transfer the raw data to this format, the batch loader goes through three steps:

- Create database – To increase the robustness of the process, the batch loader ascertains if the database has been previously created. If not, it is formed using the database name supplied from the main task. This element of the process uses the `_Data_checkDatabase` method from the `database` class contained within the `_Data_uploadCSV` method.
- Create table - Before adding the defined table, this is also checked. Using the MySQL query:

```
"SELECT COUNT( * ) FROM information_schema.tables WHERE
table_name = %s" % (convTable)
```

where convTable is the table to added, a positive or negative result can be gained. Based upon this, a new table is created. The table names are required to be added to the table. To allow this to be an automated process, a new table is created and then altered using column names extracted from the CSV file. It is assumed that the first column of the supplied CSV file contains the column names. Python supplies a CSV reading library that allows for the extraction of data from CSV files line by line. This is used to gain the necessary information.

- Load data – As the supplied raw data is in a CSV file format, MySQL supplied a process for uploading the data directly through a single statement. The data is loaded into the constructed table using a `LOAD DATA LOCAL INFILE` statement. To incorporate the use of column names extracted from the raw data, the statement is generated automatically within the `_Data_uploadCSV` method.

2.2.2 University of St. Andrews

The University of St. Andrews raw data is provided as multiple CSV files. Each file holds a separate element of a single record. Again keeping in line with previous investigation, each of the CSV files is converted to a single table before being imported to the LIDO format.

The uploading of the University of St. Andrews data follows a similar process to the Library of Birmingham. Separation from this process only occurs due to the multiple nature of the file structure.

2.2.3 Victoria and Albert Museum

The Victoria and Albert Museum raw data is provided as a single Extensible Markup Language (XML) file. The complexity of the Victoria and Albert structure, coupled with the use of XML required the use of a different approach to the uploading of both the Library of Birmingham and University of St. Andrews data. To gain access to the XML structure, the Python programming language contains an ElementTree method. This allows for the storing of hierarchical data structures in memory. XML can be represented in this way, allowing for the reading of elements and sub-elements within the tree.

The process to upload the Victoria and Albert Museum data is:

f) Create the database – A method (`Data_checkDatabase`) is used to check if the database has been added previously.

g) The XML file is checked for elements that need to be expanded, those elements that have any sub-elements.

h) Checks are required to increase the automation and robustness of the approach. If the element has sub-elements, they are navigated to. Where '_' is found as a sub-element, the previous element is used as the table name. If '_' is the element and no text is contained within the sub-element, no table is formed.

i) If a table relating to the element, or sub-element has not been created, a new table is formed. A special case is made for the element sys_id. This table is formed with the addition of a primary key of `id`. All of the tables can be linked to this table.

j) Data is added to the table created based on the information in each element. This is carried out using the `Data_addData` method.

k) Large quantities of insertions into a MySQL database can reduces the speed of its execution. To optimise the process, the transactions are committed to the database after a counter reaches 1000, reducing the overhead required.

2.3 Backing Up the Database

To mitigate the changes that the batch loader introduces to the database, a back up is made of the database to a local directory (/backup/). The back up uses the standard `mysqldump` command. If there is a need to return the database to its previous state, a SQL command can be used combined with the file produced to return the database.

2.4 Mapping to LIDO

Following the uploading of the data to the MySQL tables, each table or set of tables is mapped to the LIDO structure. Each institution has a separate configuration based on the structure of the data and the mapping of fields between the raw data and the final LIDO format. The following sections will give a brief overview of the transfer of the data from the MySQL interim tables to the LIDO format.

2.4.1 Deletion of Previous Data

In order to add the new data to the database, all previous data relating to each of the institutions is removed from the database. To achieve this, the `internal_remark` column is used. This is outside of the LIDO schema but was added to allow an ease of navigation when dealing with each separate institution.

To facilitate the removal of the data, it is necessary to disable the use of foreign key checks. This is achieved using the MySQL statement:

```
SET FOREIGN_KEY_CHECKS=0;
```

Following this command, the data from each individual institution is deleted. Once this has been carried out, the foreign key checks are reactivated. To achieve this, the below statement is used:

```
SET FOREIGN_KEY_CHECKS=1;
```

2.4.2 Mapping the Library of Birmingham Data to LIDO

The basis for each of the LIDO configuration schemas was derived from the work carried out in the cleaning, processing and formation of the initial LIDO data structures. This can be referred to in the report for Work Package 3. Python, through the use of MySQLdb, affords the use of direct interaction with MySQL databases. This approach was used in the previous steps.

Each of the mappings is carried out using SQL commands. The following is a summary of the process:

4. Insert into the fuzzyphot.lido_actor(name) the Creator from the Birmingham data

```
INSERT INTO fuzzyphoto.lido_actor (name) SELECT Creator
FROM  birmingham
```

5. Insert the photograph fields.

```
INSERT INTO fuzzyphoto.lido_photography (title,
object_description, object_published_id)  SELECT Title,
Decscription, Reference FROM birmingham
```

6. Gain the LIDO record to insert into the event table (lidoId).

```
SELECT lido_record_id FROM fuzzyphoto.lido_photography
```

7. Gain the items to insert for event.

```
SELECT Date FROM birmingham
```

8. The date is cleaned and split into a separate earliest and latest category using a dateHandler class. The date handler uses a combination of regular expressions and predefined rules to convert the data into a day-month-year format.
9. Gain the items to insert for format (formating).

```
SELECT Format FROM birmingham
```

10. Gain the items to insert for dimensions (dimensions).

```
SELECT Dimensions FROM birmingham
```

11. Insert each of the elements into the fuzzyphoto.lido_event table

```
  INSERT INTO fuzzyphoto.lido_event
  (lido_photography_lido_record_id, earliest_date,
latest_date,  material, object_measurement, event_type)
VALUES (%s, %s, %s, %s, %s, 'Shoot')" % (lidoId, earliest,
latest, formating,  dimensions)
```

2.4.3 Mapping the St. Andrews Data to LIDO

The insertion of data into the LIDO schema from the University of St. Andrews tables is relatively simplistic. The process constitutes:

5. Insert information regarding the artist into the fuzzyphot.lido_actor table.

```
INSERT INTO fuzzyphoto.lido_actor (name, birth_date,
death_date) select NamCitedName,BioBirthDate, BioDeathDate
from ColArtis
```

6. Insert the data relating to the photograph into the fuzzyphoto.lido_photography table.

```
INSERT INTO fuzzyphoto.lido_photography
(object_published_id,  object_type, title, object_description)
SELECT  ColObjectNumber, PhoRecordLevel, ColMainTitle,
ImaDescription   from ecatalog
```

2.4.4 Mapping the Victoria and Albert Data to LIDO

The structure of the Victoria and Albert museum data requires a more a more complex use of SQL statements to convert the data. The larger quantity of tables and the spread of information produces a need for the use of a SQL joins. The process is as follows:

6. Insert data into the fuzzyphoto.lido_photography table.

```
INSERT                                                INTO
fuzzyphoto.lido_photography(object_published_id,
   object_description,   link_resource,   title)   SELECT
   vanda_museum_number.vanda_museum_number,
   spec_physical_description.spec_physical_description,
url.url,   spec_title_field  FROM  vanda_museum_number  INNER
JOIN  spec_physical_description ON vanda_museum_number.sysid =
   spec_physical_description.sysid  INNER  JOIN  url  ON
```

```
vanda_museum_number.sysid = url.sysid INNER JOIN
spec_title_field ON vanda_museum_number.sysid =
spec_title_field.sysid WHERE vanda_museum_number.sysid =
(SELECT vanda_museum_number.sysid FROM vanda_museum_number)
```

7. Get the LIDO record number to insert defined as `LidoId`.

```
SELECT lido_record_id FROM fuzzyphoto.lido_photography
```

8. Get the width dimension of the photograph.

```
SELECT dimension_Height FROM dimension_Height WHERE
dimension_Height.sysid = (SELECT vanda_museum_number.sysid
FROM vanda_museum_number.sysid)
```

9. Get the height dimension of the photograph.

```
SELECT dimension_Width FROM dimension_Width WHERE
dimension_Width.sysid = (SELECT vanda_museum_number.sysid
FROM vanda_museum_number ORDER BY
vanda_museum_number.sysid)
```

10. Concatenate the width and height into a single field defined as `measurement`.
11. Insert the data into the fuzzyphot.lido_event table.

```
INSERT INTO fuzzyphoto.lido_event(material, earliest_date,
latest_date, object_measurement, technique,
lido_photography_lido_record_id, event_type) SELECT

mus_materials_techniques_note.mus_materials_techniques_note,
earliest.earliest, latest.latest, %s,
spec_title_field.spec_title_field, %s, 'Shoot' FROM
mus_materials_techniques_note INNER JOIN earliest ON
mus_materials_techniques_note.sysid = earliest.sysid INNER
JOIN latest ON mus_materials_techniques_note.sysid =
latest.sysid INNER JOIN spec_title_field ON
mus_materials_techniques_note.sysid =
spec_title_field.sysid WHERE
mus_materials_techniques_note.sysid = (SELECT
vanda_museum_number.sysid FROM vanda_museum_number)%
(measurement, lidoId)
```

3. Conclusion

This report outlines the structure of the batch loader to be used in the Fuzzyphoto project. It supplies an insight into the operation of the process to maintain sustainability of the data for the construction of similarity links between meta-data relating to historical images.

4. Appendix

To run the included Python files, Python must be installed. The Python files are designed for a Linux based system. Most modern versions of Linux come with Python out of the box such as Ubuntu, Fedora, Redhat Enterprise (RHEL) and CentOS.

If the user is within the PartnerData structure, the python script can be run as:

```
python mainProcess.py -t <timestampFileLocation>
-u <uploadedFilesDirectory>
```

Running the batch loader script from the terminal requires two additional arguments. If theses are not passed, the default parameters will be used. The default are located as an relative `../upload/` directory within the PartnerData structure. Within the upload folder, the `timestamp_config.csv` file will be contained. This file holds the timestamps relating to the files uploaded previously.

Appendix 10. The FuzzyPhoto MySQL Cluster

The FuzzyPhoto MySQL Cluster

Table of Contents

1. Synopsis

This document has been constructed to assist users of the MySQL cluster established for the purposes of the FuzzyPhoto project. The cluster is hosted in the Portland Building within De Montfort University. The document will assume some previous experience with Information Technology (I.T) and SQL commands. It will cover:

- Background to the purpose of the MySQL cluster.
- The hardware structure of the MySQL cluster.
- The software structure of the MySQL cluster.
- The process of initiating the cluster.
- The process of restarting the cluster.
- How to migrate a database onto the cluster from a standard MySQL export.

2. Background

The FuzzyPhoto MySQL cluster serves as the main data repository for the FuzzyPhoto project. Within the cluster a composition of data from the Exhibitions of the Royal Society 1870 – 1915 (stored at De Montfort University) and partner organisations is stored. The data is composed of meta-data relating to historical photographs within these collections. The cluster also serves to assist in the processing of links between these collections. The structure of the cluster was to maintain data isolation from external access along with allowing continued, sustained expansion of core structure. Figure 1 shows how the cluster relates to other elements within the process. The main output from the cluster is a links database. Users can access this database via a web server (more detail is given in additional documents). The links database is periodically updated from the cluster through a defined link.

3. Hardware Structure

The FuzzyPhoto is constructed across three Dell PowerEdge R320 servers. Each server has the below specification:

- CPU – Xeon E5-2403 1.80GHz, 10M Cache.
- Network – Broadcom 5720 Dual port 1GB ethernet.
- Memory – 4GB UDIMM, 1333MHz.
- Hard Drive – 2 x 500GB SATA, RAID 1 configured.

Each server runs a separate installation of the Ubuntu 12.04 LTS Precise Pangolin server. The hardware structure is shown in Figure 1.

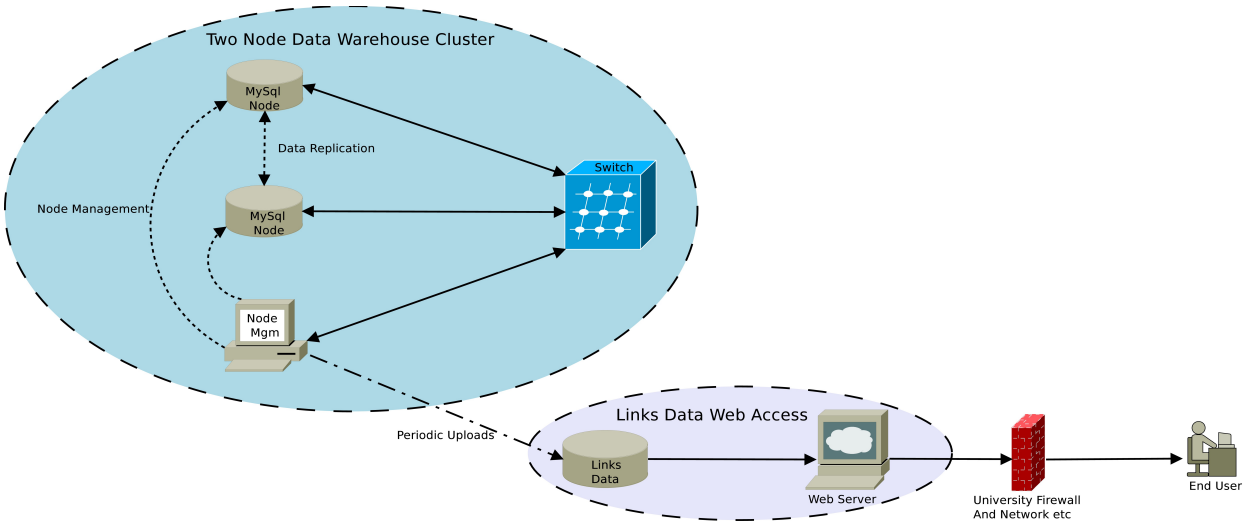


Figure 1. Structure of the FuzzyPhoto Cluster.

The cluster is composed of three physical devices. The three servers house separate components of the cluster structure alongside a web server used within the FuzzyPhoto project. The web server is isolated from the cluster through both software and hardware interfaces. Each of the three servers are linked via a gigabit switch. Although the web server is hosted on one of the servers, it is not physically linked to the cluster network.

The cluster is controlled via a KVM switch. This allows access from a single terminal. In position 1, the switch relates to the web server / manager, position 2 it relates to cluster slave 1 and position 3, it relates to cluster slave 2.

3. Software Structure

The cluster is composed of three separate elements: Management, Data and SQL Nodes.

3.1 Management Node

The MySQL management node resides on kmd4. Unlike the other devices it is a virtual machine hosted on the FuzzyPhoto web server. This allows isolation of the system from user access. The management node is also isolated through separate network access. In MySQL cluster, there is no single central server or process. All the agents that are involved in the process collaborate in managing the cluster as a whole. The management node is responsible for:

- Starting, stopping, and restarting cluster nodes.
- Cluster configuration.
- Cluster software upgrades.
- Host and node status reporting.
- Recovery of failed nodes.

3.2 Data Node

The data nodes are used to store the data within the cluster structure. Tables are automatically sharded (horizontal partitioning of data) across the data nodes which also manage load balancing, replication, failover and self-healing.

3.3 SQL Node

The SQL nodes act in a similar manner to standard SQL databases. They are MySQL servers that connect the data nodes in order to perform storage and retrieval of information.

4. Initiating Cluster

The process of initialising and starting the cluster needs to be carried out in a set order. This is:

11. Start the management node.
12. Initialise the data nodes.
13. Connect the SQL nodes.

This process will be described taking each step in turn highlighting the necessary command-line statements to use. The command-line statements will be given as

```
cd /usr/local/
```

To enter a command, first open a terminal. This can be accomplished by pressing *ctrl + alt + t.*

4.1 Start the Management Node

Switch the KVM to position 1. This will allow the keyboard and mouse to connect to KMD4. Within the KMD4 operating system is contained in a Virtual Machine (VM). A VM is a software based operating system that emulates the functions of a real world computer architecture. The VM is run using Virtual Box. The management node VM can be started from the command-line:

```
vboxmanage startvm clusterManager
```

where `clusterManager` is the VM being started.

4.1.1 Locate the Correct Directory

Using the command-line, navigate to the location of the management node.

```
cd /usr/local/mysql
```

4.1.2 Initialise the Management Server

In the terminal, run the following command:

```
./bin/ndb_mgmd -f configini --initial
--configdir=/var/lib/mysql-cluster #Note: all one line
```

The ndb_mgmd calls the cluster management server. The `-f` and `--configdir` options call a configuration file for the server from the defined location. Using the `--initial` option forces the server to read from the configuration file each time it is loaded rather than reading the global configuration.

4.1.3 Start the Management Client

In the terminal, run the following command:

```
sudo ./bin/ndb_mgm
```

This command must be run with root privileges. This will start the management client process. This is not required to mange the cluster but is used to monitor the processes.

To view the connections to the manager, enter

```
SHOW
```

At this point there will be no connections shown.

4.2 Start the Data and SQL Nodes

On each of the additional servers there resides both a data node and a SQL node. The following commands need to be carried out on both servers to start the nodes. Use KVM switch position 2 and 3 to operate the input devices on cluster slaves 1 and 2 that house both data and SQL nodes.

4.2.1 Start the Data Node

Using the command-line, navigate to the location of the data and SQL nodes.

```
cd /usr/local/mysql
```

In the terminal, run the following command:

```
sudo ./bin/ndbd --initial
```

This command must be run with root privileges. This will restart the data node using an initial restart process.

Following on, enter this command:

```
sudo ./bin/mysqld --user=root &
```

This command must be run with root privileges. This will restart the data node using an initial restart process.

4.3 Managing the Connections

The management client can be used to view and manage the connections to the cluster. This is started by running the commands in Section 4.1.3. Below are a set of commands that can be run to assist in managing the cluster.

l) `HELP` *– displays information about all available commands.*

m) `SHOW` *– Displays information on the clusters status. Use this after starting the data and SQL nodes to check the connections have been started successfully. Possible node statuses include* `UNKNOWN, NO_CONTACT, NOT_STARTED, STARTING, STARTED, SHUTTING_DOWN and RESTARTING.`

n) `node_id START` *– This brings online a specified node.*

o) `ALL START` *– Brings all nodes online except management nodes.*

p) `node_id STOP` *– Stops the specified management or data nodes*

q) `node_id RESTART` *– Restarts the specified node. Options can be used including* `-i` *for an initial restart.*

r) `node_id STATUS` *– Gives the status of the specified node.*

s) `MemoryUsage` *– Gives the memory usage and index memory being used by each data node*

t) `SHUTDOWN` *– Shuts down all the cluster data nodes and management nodes.*

u) `EXIT, QUIT` – Terminates the management client.

5. Importing and Migrating MySQL databases from innoDB or MyISAM to NDB

For replication to occur across the cluster, a different engine (type of table) to those often used on a desktop or server configuration needs to be used. For a table to be replicated the engine needs to be defined as `NDBCLUSTER` or `NDB`. A table defined in either `innoDB` or `MyISAM` can be added to the cluster but it will not be replicated. The following process will describe how an SQL dump can be imported and converted to the required engine format. A MySQL dump is a program that can be used to dump a database or a collection of databases for backup or transfer to another SQL server. The dump contains SQL statements to create and/or populate the required tables.

5.1 Import Database to SQL nodes

To import a SQL dump initially there maybe a requirement to uncompress the file. On each data node, uncompress the files in an accessible directory. Depending on the format, this can be achieved using the below statements.
*If the file is *.tar*

```
tar xvf archive_name.tar
```

*If the file is *.tar.gz*

```
tar xvfz archive_name.tar.gz
```

*If the file is *.tar.bz2*

```
tar xvfj archive_name.tar.bz2
```

Once the file(s) have been uncompressed, they can be imported directly onto the cluster SQL servers. The first step in the process is to create a database on the server to house the tables. Firstly, connect to the server using the below command.

```
./mysql – h 127.0.0.1 –P 3306
```

The h option is the host (local in this instance) and P is the port. This is the defined port for the server.
Next a database is created using the below command, followed by exiting the SQL command-line.

```
CREATE DATABASE database_name;

EXIT;
```

The tables from the uncompressed SQL dump can be imported using the following command where `database_name` is the database created and `tables.sql` is the uncompressed SQL dump.

```
sudo ./mysql – h 127.0.0.1 –P 3306 database_name < tables.sql
```

Once the database has been updated, the tables engine needs to be converted. This can be achieved by re-connecting to the SQL server.

```
./mysql – h 127.0.0.1 –P 3306
```

The following SQL command outputs a list of SQL commands that can be copied to the command-line to alter all of the tables within the database.

```
SELECT CONCAT ('ALTER TABLE', table_name, ' ENGINE=NDB;') AS
sql_statements FROM information_schema.tables AS tb WHERE
table_schema = database_name ORDER BY table_name DESC;
```

Within the command database_name needs to be replaced by the database being

used. This will return a list of SQL statements. These can be copied into the SQL server command-line to convert all of the tables within the database. To assist in the copy and pasting of the commands into the command-line, copy the statements into the gedit text editor. Replace all "\|" with "". This will remove all unneeded characters. The resulting statements can be copied directly into the command-line.
If a single table needs to be converted, this can be achieved using the following command.

```
ALTER some_table ENGINE=NDB;
```

This will change the engine of `some_table` *to the required engine so that the cluster will replicate it across all nodes.*

Appendix 11. WP 5 Word Sense Disambiguation report

FuzzyPhoto AHRC AH/J004367/1

Work Package 5 Report: Word Sense Disambiguation

Dr. Jethro Shell, Dr. Simon Coupland and Prof. Stephen Brown

August 21, 2013

Status: 2nd draft

Distribution Type: Internal

Keywords: Word Sense Disambiguation, Natural Language Processing, Part-of-Speech Tagging, Fuzzy Logic

Contents

1. Introduction

Incomplete and vague data are often common in art and humanities research, particularly in the data that is associated with historical artefacts such as photographs. Many collections suffer from inadequate documentation with researchers relying on the knowledge of curators to often identify relevant research materials. The increasing availability of resources through digital access and powerful search tools has opened up the opportunities to discover these resources, however the structure of the data can limit the accessibility.

The FuzzyPhoto project looks to investigate the use of Computational Intelligence (CI) techniques to identify relationships within the data through the production of a *finding aid* to suggest likely matches of photographs across different historical collections.

The FuzzyPhoto project aims to exploit information held within photographic metadata (data about the photographic data) contained within an AHRC funded database held by De Montfort University, the Exhibitions of the Royal Photographic Society 1870-1915 (ERPS) along side a number of partner databases. Photographic records provide a unique problem as they can be exhibited/published multiple times on different occasions using different titles by different people. Assigning a title to a specific image, as a result, can be complex.

In comparing metadata of the photographs held, a semantic similarity measure is used. A component of the similarity process is the use of query expansion. Query expansion is the process of using similar meanings to those in a query to increase the chances of matching words within the query itself [Xu et al. (1996)]. To undertake query expansion, an understanding of a word is required. The context surrounding individual words can influence the meaning. Understanding a word in its context is seemingly easy for the humans but a complex and difficult task for computers, the word *fast* can mean *rapid* but also *motionless* dependent on the context. To understand the context of words in a sentence, Part-Of-Speech Tagging (POST) can be used. POST is the application of descriptors for each element in a sentence to help disambiguate the words, for example the structure of the tagging output can be represented as:

The/DT Eiffel/NNP Tower/NNP looked/VBD beautiful/JJ in/IN the/DT sun/NN

where DT is a determinant, NNP is a proper noun (singular), JJ is an adjective, IN is a preposition, VBD is a verb (in the past tense) and NN is a noun (singular).

This report details the research carried out into the application of POST to assist Word Sense Disambiguation (WSD). POST was applied to photographic titles within the ERPS database with the aim of achieving an automated tagging process. The automated disambiguation of titles was highlighted as a required component in the FuzzyPhoto record matching process. To achieve this, a proposed method combining the use of fuzzy logic and probability was investigated. This was compared to two available software Part-of-Speech taggers. A comparison was made based upon accuracy. It was found that the fuzzy approach was overall less accurate than the compared software when used against a test dataset. Tagging 100 titles from the ERPS database consisting of 465 words, the fuzzy methodology achieved 83.65% accuracy. The Natural Language Toolkit (NLTK) tagger, in comparison, achieved 85.80%, with the Stanford software producing 86.23%. The fuzzy approach was more accurate at identifying particular word types within the structure of the sentence than the other methods compared. Overall, a recommendation was made to use the Stanford tagger within this project.

Outputs of this workpackage comprise:

14. This report.
15. Part-Of-Speech tagged annotated records of the ERPS database.
16. A Java implementation of Fuzzy Part-Of-Speech Tagging.

The elapsed time for this work package was 7 months. The resources required to complete this work package were 84 person-days.

2. Word sense disambiguation

Word Sense Disambiguation (WSD) is an open problem within computational linguistics. Understanding the meaning of a word in its context is something that is familiar and seemingly easy for a human, yet complex to a computer. For example, the word light can mean not heavy or illumination. The sentence *He turned on the light* is clear as to the meaning although the process to decide this is complex. Stevenson and Wilks defined WSD as

> "the process of identifying the meaning of words in context [Stevenson and Wilks (2003)] "

The FuzzyPhoto project uses elements of meta-data that are descriptive. They contain sentence structure that require contextual understanding. One of the steps within the FuzzyPhoto approach is the use of query expansion. Query expansion is the process of using similar meanings to those in a query to increase the chances of matching words within the query itself [Xu et al. (1996)]. To undertake query expansion, an understanding of a word is required. This is referred to as disambiguation.

There are a number of strategies that can be employed to disambiguate words within a context. Many focus on the use of a corpus taken from a specific subject area. The FuzzyPhoto project is different in this manner. There is not a single subject that encompasses the collections, as with medical documents or items from a specific location. Additionally the structure of the titles that are being focussed on are extremely sparse. Standard disambiguation approaches use text structures that encompass long sentences, or even paragraphs. This allows a deeper understanding of the context. The photographic metdata holds short text, the title averaging only 6.1 words. This produces a more difficult problem domain.

One element of WSD is the use of Part-Of-Speech Tagging (POST). POST is the process of assigning descriptors, or tags to input tokens within a sentence structure. The application of word classes has been used within linguistics since c.100 BC when Thrax used classification words such as noun, verb, participle, article, pronoun, preposition, adverb, and conjunction [Voutilainen (2003)]. POST is predominantly used as a preprocessing measure. It can be carried out quickly and more accurately than parsing, and development of taggers for specific domains is more rapid [Brill (2000)] . Specific POST systems have been focussed on individual problem domains such as Twitter [Gimpel et al. (2010)] and across languages [Snyder et al. (2009)].

The output from a POST process can be used to further disambiguate a sentence structure or be directly used as the sole disambiguation. The structure of the tagging output can be represented as:

> The/DT Eiffel/NNP Tower/NNP looked/VBD beautiful/JJ in/IN the/DT sun/NN

where DT is a determinant, NNP is a proper noun (singular), JJ is an adjective, IN is a preposition, VBD is a verb (in the past tense) and NN is a noun (singular).

In this report, the use of a combined Fuzzy Logic (FL) and probabilistic method will be discussed. The focus of the method will be the extraction of word tags for 100 test titles from the Exhibitions of the Royal Photographic Society (1870-1915) (ERPS) database. The method will be compared to two developed, open source software applications based upon the accuracy of the tagging process.

3. Comparative Work

There are a large number of systems that have implemented POST. In this section a brief summary will be given of some of the approaches and current systems that are available.

3.1 Bidirectional Processing

Many processes approach the problem of POST t_1 as a unidirectional sequence problem. Despite the application of different algorithms, the system will navigate through the sequence in a single direction, left to right or right to left. In a standard left to right first-order Hidden Markov Model (HMM), the current tag t_0 is predicted based on the nature of the preceding tag,[Toutanova et al. (2003)]. Despite unidirectional models ability to capture the information of both directions of the model, as this is implicit when the next word is generated t_{-1} , bidirectional models have been proposed. Toutanova et al. [Toutanova et al. (2003)] exploit the use of dependency networks to efficiently infer more information from the sequence using both directions. Additionally they incorporated multiword feature templates so that idiomatic word sequences could be learnt and used within the model. Based on the model, Toutanova et al. were able to produce an accuracy of 97.24% per tag and 56.34% on correct whole sentence identification for the Penn Treebank WSJ dataset [Marcus et al. (1993)].

The work of Toutanova et al. [Toutanova et al. (2003)] forms the basis for the Stanford log-linear part-of-speech tagger [11]. The software implementation will be used as a comparison to the methodology in this report.

3.2 Support Vector Machine

Within classification problems, the use of a Support Vector Machine (SVM) is a popular approach. Giménez and Màrquez [Giménez and Màrquez (2004)] discuss the use of a SVM for POST. A SVM is a process to classify data by maximising the solution between two groups. Figure 1 shows the optimal hyperplane that is produced between two groups by maximising the margin at the closest points. The SVM is predominantly used to form a model of the POST domain based upon training data. The model can then be applied to the classification problem. Giménez and Màrquez used a SVM approach to tag 18 sections of the Wall Street Journal corpus. 2.81% of the words in the dataset were unknown to the model. They were able to generate tags with between 96.89% and

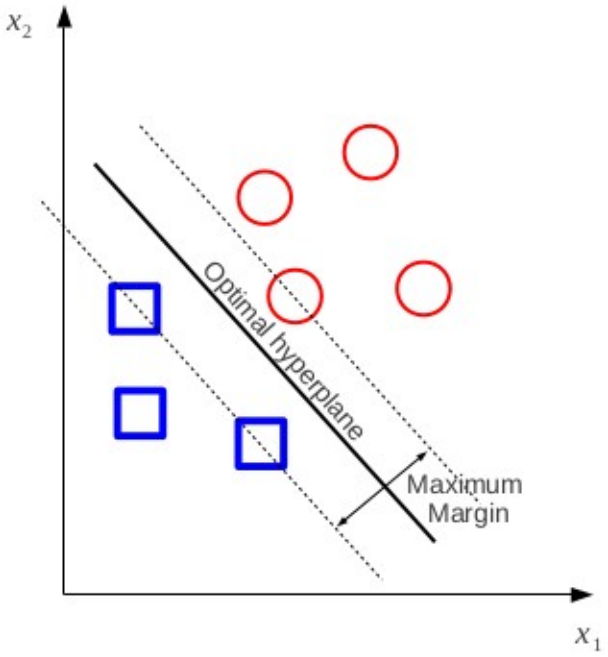


Figure 1. Outline of Support Vector Machine

98.96% accuracy.

3.3 Hidden Markov Model

There are a number of implementations that use a HMM [Manning (2011)], [Gimpel et al. (2010)]. In principle a HMM is a tool for representing probability distributions over a sequence of observations. A HMM is highly applicable to POST as a constituent part of the Markov process is to encapsulate all of the history of a process to predict the future of that process. Cutting et al. [Cutting et al. (1992)] used a HMM method to approach POST. Their strategy was to initially generate a HMM based on annotated sentences. The model was tuned using empirical and *a priori* information. The model was trained on approx. 500,000 words from the Brown corpus [Marcus et al. (1993)]. When applied to an equal quantity, a 96% accuracy was gained.

4. FuzzyPhoto Approach

This report proposes a new strategy for POST using a combination of probability and fuzzy sets to represent the uncertainty when defining a word type. This can be categorised as fuzzistics, the merging of the words fuzzy and statistics. Originally used in [Mendel (2007)], fuzzistics describes the problem of going from word data collected from a group of subjects quantified by statistics, with the uncertainties that are inherent, to a word fuzzy set model that captures the uncertainties within the word data [Gimpel et al. (2010)]. The outlined process uses a group of probabilistic methods to produce a group of differing values relating to each possible word type for each word within the title. These groups are

captured within type-1 fuzzy sets. The application proposed incorporates three sources of information. The sources of data come from the use of the Stanford parser [Toutanova et al. (2000)], an annotated corpus using data from the ERPS database and data captured from the WordNet lexical database [Miller and others (1995)].

The methodology is composed of four distinct stages:

- Language Recognition.

- Word Type Probability.
- Fuzzification
- Set Comparison

Each of these stages will be addressed in sequence using a simple example. To demonstrate the application of the approach, the use of a two word title will be used: **daffodil fair**.

4.1 Stage One: Language Recognition

The overall probabilistic structure is based upon the composition of the individual probabilities that a word embodies a particular type. Each word is contextualised by the words that exist around it. The words that precede a word will influence the word type. The titles within the ERPS database contain multiple language types. To overcome this issue, two third-party processes were used to identify and translate non-English language words. The process used a combination of information from language detection software [Shuyo (2010)] and the Microsoft Bing web-based translator. A reduced set of languages were selected within the detection software. It was found that predominantly French, German, Spanish and Dutch were within the ERPS dataset.

4.2 Stage Two: Word Type Probability

The second stage of the process produces a series of probabilities relating word types to each word within the title. The POST word types used are mapped to those within a modified WordNet structure. These types were: adjective, adverb, article, conjunction, determiner, noun, other, preposition, pronoun, and verb. There are three separate information sources that provide probabilities of the word formats. Each of these will be taken in turn.

1. Parsed Title - Extraction of the probability of POST through the use of the Stanford tagging software.
2. Annotated Text - Application of sentence sequence structure based on an annotated corpus.

- WordNet Frequency Count - Incorporation of word type frequency counts using the WordNet corpus.

4.2.1 Parsed Title

Each of the titles contained within the corpus are processed through the Stanford Parser [Toutanova et al. (2000)]. The parser software is available as a downloadable library (see http://nlp.stanford.edu/software/lex-parser.shtml) that can be implemented through the use of the Java programming language. Additionally there are command line options and currently an online version. The simplest approach to produce an output is to use the command line operation. There are a number of options built into the parser that can be refined. A sentence can be parsed using:

```
java -mx200m -cp "stanford-parser.jar:." ParserDemo2 englishPCFG.ser.gz testsent.txt
```

where testsent.txt is the text file. The sentence is passed through the parser as a whole. The parser is configured to supply a normalised probabilistic value of the tag for each word based on the whole sentence. Taking our toy example daffodil fair, the parser returns a value for daffodil of 0.8141 for the tag noun. In the Fuzzy approach, a Java based implementation was used. The library was configured to incorporate the standard englishPCFG training structure.

4.2.2 Annotated Text

To supplement the information gained from the Stanford parser, an annotated corpus based on the ERPS database was used. Each record was taken and tagged by members of the FuzzyPhoto team experienced in data analysis. Labelled text is predominantly used to train systems to directly tag items. The sparsity of the available text coupled with the depth of subject area made this a difficult task. To gain information from the available data, the tag structure was extracted from 100 labelled titles taken from the ERPS database. This was based on a left-to-right sequence with a bi-gram structure. A probability of a second word type occurring was produced based upon the preceding word type. At the start of the sentence, the first word would be defined without the use of this process.

Taking the previously defined example, Daffodil is defined as being a noun. Based upon the accrued bi-gram information, the following word is assigned a set of tags mapped to probabilities. The top three probability values can be seen in Table 1.

W_1	Tag_1	W_2	Tag_2	Probability
daffodil	noun	fair	article	0.0283
daffodil	noun	fair	noun	0.2291
daffodil	noun	fair	adjective	0.0283

Table 1: Bigram of Word Sequence

4.2.3 WordNet Frequency Count

The third stage uses the WordNet software structure [Fellbaum (1998)]. At the heart of WordNet is an annotated corpus that contains over 117,000 synsets of nouns, verbs, adjectives and adverbs. WordNet contains a frequency of the occurrence of each annotated word within each synset group. Based upon this frequency, a probability of each word within the sentence structure is generated.

Using the Daffodil Fair example, WordNet would output the below probabilities:

$$daffodil \rightarrow noun \rightarrow 1.4634^{-4}$$
$$fair \rightarrow adjective \rightarrow 0.0196$$
$$fair \rightarrow noun \rightarrow 1.4634^{-4}$$
$$fair \rightarrow verb \rightarrow 9.309210^{-5}$$

Based on the frequency values, the tags noun and adjective would be used. Other values produced an output of 0 as no frequencies were defined.

4.3 Stage Three: Fuzzification

The values from each of the three component groups are fuzzified into a set representation. This uses the concepts set out by Lotfi Zadeh [Zadeh (1988),Zadeh (1965) of fuzzy logic. Fuzzy logic is an approach to representing degrees of truth rather than the classical view of logic as only true or false. A fuzzy set can be used to assign an element a degree of membership rather than simply a boolean value. This can represent *absolutely true* and *absolutely* false, but equally any degree in between. This assists in representing cases that are vague such as propositions like *this person is old.* Fuzzy sets are constructed in a domain that represents the subject matter. In the case of the implementation demonstrated here, the sets are formed based on the domains from each information group, and from each word type. Each set, constructed as a triangular set here though this can be a

number of different functions, represents the linguistic values of Very Small (VS), Small (S), Medium (M), Large (L) and Very Large (VL). The sets illustrate the probability that the word *daffodil* is a noun. Figure 1 shows that *daffodil* produces a membership value in the set of Large of 0.8. It also produces a membership value of 0.2 in the set of Very Large.

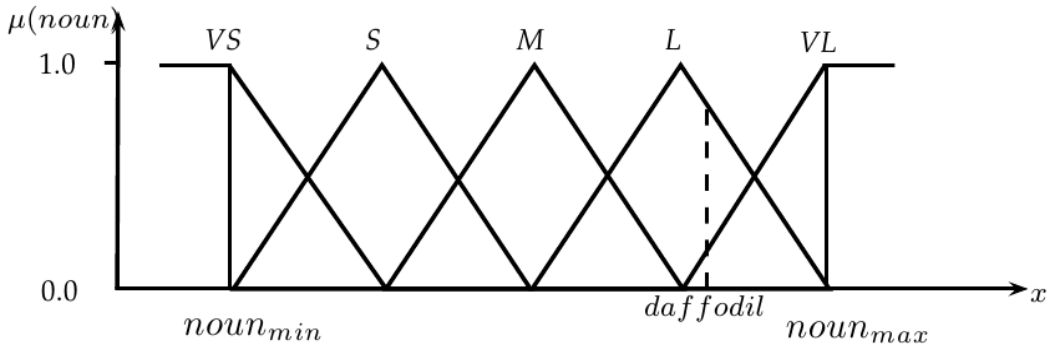


Figure 2. Fuzzy Representation of the Probability of the Word *daffodil* Produced By the Stanford Parser.

Once each of the component probabilistic elements are represented as fuzzy sets, each value is combined as a Fuzzy Weighted Average (FWA). The FWA was first proposed by Dong and Wong [Dong and Wong (1987)]. There have been a number of variations based on this original version [Lee and Park (1997),Liou and Wang (1992),Wu and Mendel (2010)]. This implementation uses the original structure. A FWA can be depicted using the following definition. If $A_{1,}A_{2,}..,A_n$ and $W_{1,}W_{2,}..W_n$ are fuzzy numbers defined in the universes $X_{1,}X_{2,}..,X_n$ and $Z_{1,}Z_{2,}..,Z_n$ then a fuzzy weighted average can be:

$$y=f\left(x_n,w_n\right)=\frac{x_1w_1+x_2w_2+...x_nw_n}{w_1+w_2+...w_n}$$

where for each $i=1,2,...,n,\ \ x_1\in X_i,w_i\in Z_i,$ and $w_1+w_2+...+w_n>0$ .The use of a weighted average allows for the methodology to exert differing influence across the component elements within the system. Fuzzy sets were constructed based upon knowledge elicited from the data by the FuzzyPhoto team. The sets consisted of input values of each of the three components: Parsed Title, Annotated Text, WordNet Frequency Count. The annotated text was given the lowest weighting. Despite being data from the same context, the relatively small quantity of information was found to skew the test data. Larger weightings were given to the parsed and WordNet information.

4.4 Stage Four: Set Comparison

To compare each of the component sets within the method, a simple set comparison method was employed. The output from the FWA is a fuzzy set. A fuzzy set is produced for each element within each of the domains. For example, the annotated data element will produce a fuzzy set for each of the word types adjective, adverb, article, conjunction, determiner, noun, other, preposition, pronoun, and verb. These are defuzzified to produce a single value using the Centre of Gravity (CoG) [Zadeh (1994)] approach. Each of the values are compared. The highest value from each of the types classifies the corresponding word. Based on the example, the fuzzy method outputs daffodil as a noun, and fair also as a noun.

5. Experimentation

To test the methodology, a dataset of 100 titles from the ERPS database were additionally annotated by members of the FuzzyPhoto team along with the 100 titles used to build the model. The 100 annotated titles acted as ground truth. The methodology was applied to these values to assess its performance.

A comparison was also made against two readily available software implementations, the Stanford Part of Speech Tagger [Toutanova et al. (2000)], and the Natural Language Toolkit (NLTK) [Bird (2006)] Python software library.

An example of the format of the output from the fuzzy approach is shown in Table 2

Title No	Word	Type	Word Annotated	Type Annotated
3	triptographic	adjective	Triptographic	noun
3	cameos	noun	Cameos	noun
4	portrait	noun	Portrait	noun
4	of	preposition	of	preposition
4	a	determiner	a	article
4	lady	noun	Lady	noun
5	the	determiner	The	article
5	village	noun	Village	noun
5	wholesale	adjective	Mayor	noun
6	the	determiner	The	article
6	old	adjective	Old	adjective
6	church	noun	Church,	noun
6	bonchurch	verb	Bonchurch	noun
7	cloudland	adjective	Cloudland	noun
7	sunset	noun	Sunset	noun
8	high	adjective	High	noun
8	rocks	noun	Rocks,	noun
8	cheddar	noun	Cheddar	noun
8	cliffs	noun	Cliffs	noun
9	changing	verb	Changing	verb
9	box	noun	Box	noun
9	for	noun	for	preposition
9	dry	noun	Dry	verb
9	plates	noun	Plates,	noun
9	and	adverb	and	conjunction
9	expanding	adjective	Expanding	verb
9	camera	noun	Camera	noun

Table 2. Comparison of Fuzzy Method to Ground Truth Data.

In order to standardise the input, all of the text was reduced to lower case. A value of one was given to each correct tag, and zero to an incorrect tag. A standard error rate was produced from this process.

Running the fuzzy method across the annotated ground truth titles produced a word by word accuracy of 83.65% across the 100 titles. The titles were composed of 465 words in total. Table 3 shows a breakdown of the results by type.

Type	Total	Identified	Correctly Identified %
adjective	33	12	36.36
adverb	12	5	41.67
article	44	44	100
conjunction	17	17	100
noun	282	248	87.94
preposition	52	49	94.23
pronoun	2	2	100
verb	23	12	52.17

Table 3. Percentage of Correctly Identified Types For the Fuzzy Method.

The methodology was efficient at tagging articles, conjunctions, prepositions, pronouns and nouns. The system incorporated two basic expert constructed rules to assist the tagging. These facilitated the tagging of conjunction and preposition words. All “the” and “a” words were

classified as prepositions. and words were defined as conjunctions. Further expansion of basic rules could bring further results. The quantity of pronouns within the test set was limited (only two). This limits the impact of this group. The system was less able to identify adverbs, adjectives and verbs. The system incorrectly identified a number of adjectives as nouns. This may result from the contextual nature of the title combined with the weighting structure. This requires further investigation.

To form a comparison, the test dataset was run against the NLTK implementation of a Part-of-Speech tagger. The implementation that was used was configured on the bi-gram tagger structure and trained using the Penn Bank annotated text. The NLTK software was able to achieve an 85.80% accuracy. The NLTK implementation outperformed the fuzzy method in the adverb (6), noun (263) and preposition (50) groups. It performed equally or less well in the other categories.

The test dataset was also processed using the Stanford tagger. The software is readily available and can be trained on any specific set of data. As no large set of annotated data was available, the tagger was trained using the supplied English language model. Overall the tagger was able to achieve an accuracy of 86.23%. It outperformed the fuzzy method within the adjective (14) and noun (258) categories, but as with the NLTK software it performed less or equally well in the other categories.

The comparison methods were more accurate in tagging the noun group. This was the largest group in the test dataset. The fuzzy method was able to outperform or match the NLTK implementation in four of the groups achieving 7% higher accuracy across those groups. In comparison to the Stanford methodology, the fuzzy method achieved equal or greater accuracy in five of the categories with a 0.02% increase.

6. Conclusion

The main conclusion of this report is to suggest that the Fuzzy Photo project adopts the use of the Stanford POST methodology. The combination of accuracy, when used on the ERPS test dataset, and speed of implementation are in tune with the developmental process.

However, there is scope to further investigate and expand the fuzzy disambiguation methodology. The use of fuzzy sets to represent the part-of-speech tags is able to assist in the representation of the uncertainty that is contained within text. The fuzzy method was close in performance to the two methods compared. Certain categories within the titles were shown to be outperformed by the fuzzy method in comparison to the NLTK and Stanford software.

Further work may yield more results. A greater corpus to extract the probabilistic values, and the use of bidirectional processing may improve the accuracy gained from the methodology. Additionally, the adoption of further rules within the system may refine the process.

References

Bird S. (2006) NLTK: the natural language toolkit. , 69-72.

Brill E. (2000) Part-of-speech tagging. *Handbook of Natural Language Processing* , 403-414.

Cutting D., Kupiec J., Pedersen J. and Sibun P. (1992) A practical part-of-speech tagger. , 133-140.

Dong W. and Wong F. (1987) Fuzzy weighted averages and implementation of the extension principle. *Fuzzy sets and systems* **21**, 183-199.

Fellbaum C. (1998) WordNet: An electronic database. , .

Giménez J. and Màrquez L. (2004) SVMTool: A general POS tagger generator based on Support Vector Machines. , .

Gimpel K., Schneider N., O'Connor B., Das D., Mills D., Eisenstein J., Heilman M., Yogatama D., Flanigan J. and Smith N. A. (2010) Part-of-speech tagging for twitter: Annotation, features, and experiments. , .

Lee D. H. and Park D. (1997) An efficient algorithm for fuzzy weighted average. *Fuzzy sets and systems* **87**, 39-

45.
Liou T.-S. and Wang M.-J. J. (1992) Fuzzy weighted average: an improved algorithm. *Fuzzy sets and systems* **49**, 307-315.
Manning C. D. (2011) *Part-of-speech tagging from 97% to 100 is it time for some linguistics?*. In: (Ed.), *Computational Linguistics and Intelligent Text Processing*, Springer.
Marcus M. P., Marcinkiewicz M. A. and Santorini B. (1993) Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics* **19**, 313-330.
Mendel J. M. (2007) Computing with words and its relationships with fuzzistics. *Information Sciences* **177**, 988-1006.
Miller G. A. and others (1995) WordNet: a lexical database for English. *Communications of the ACM* **38**, 39-41.
Shuyo N. (2010) Language Detection Library for Java. , .
Snyder B., Naseem T., Eisenstein J. and Barzilay R. (2009) Adding more languages improves unsupervised multilingual part-of-speech tagging: a Bayesian non-parametric approach. , 83-91.
Stevenson M. and Wilks Y. (2003) Word sense disambiguation. *The Oxford Handbook of Comp. Linguistics* , 249-265.
Toutanova K., Klein D., Manning C., Morgan W., Rafferty A., Galley M. and Bauer J. (2000) Stanford log-linear part-of-speech tagger. , .
Toutanova K., Klein D., Manning C. D. and Singer Y. (2003) Feature-rich part-of-speech tagging with a cyclic dependency network. , 173-180.
Voutilainen A. (2003) Part-of-speech tagging. *The Oxford handbook of computational linguistics* , 219-232.
Wu D. and Mendel J. M. (2010) Ordered fuzzy weighted averages and ordered linguistic weighted averages. , 1-7.
Xu, J., & Croft, W. B. (1996) Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 4-11). ACM.
Zadeh L. (1965) Fuzzy sets*. *Information and control* **8**, 338-353.
Zadeh L. (1988) Fuzzy logic. *Computer* **21**, 83-93.
Zadeh, L. (1994) Soft computing and fuzzy logic *Software, IEEE*, **1994**, *11*, 48 -56

Appendix 12. WP 6 FuzzyPhoto widget implications for partners

What is the widget?

The FuzzyPhoto widget is a small piece of code that can be inserted into partners' web pages to display links to corresponding objects in each other's catalogues. Visitors to an object record on a webpage will see a list of hyperlinks suggesting possible matches, with varying degrees of confidence. Following one of these links will open a new window containing the suggested matching object record within its owner's web site.

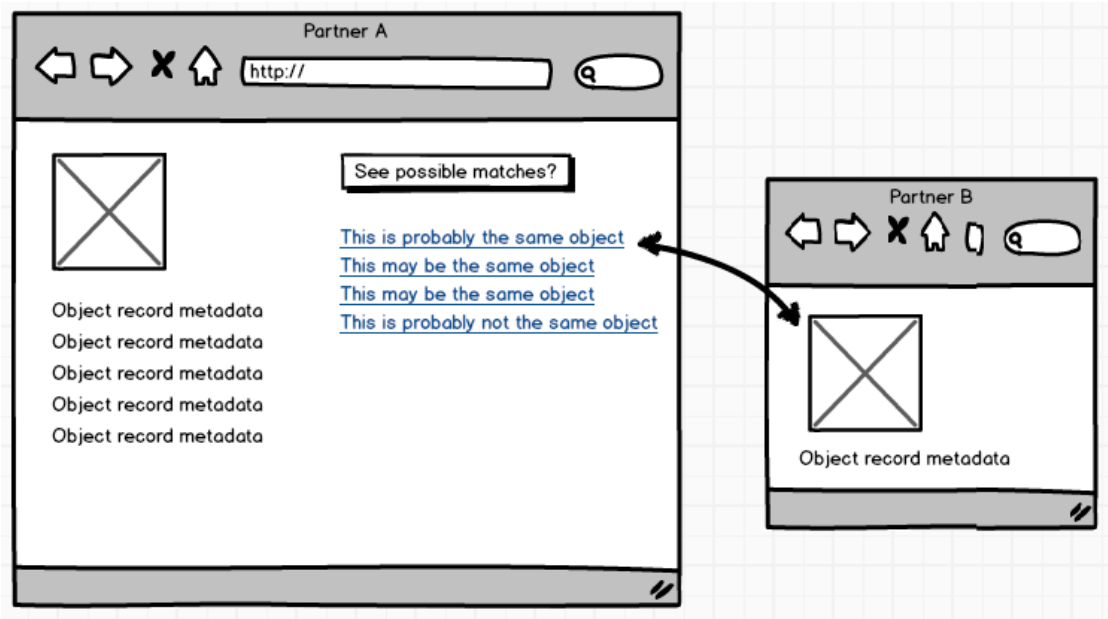


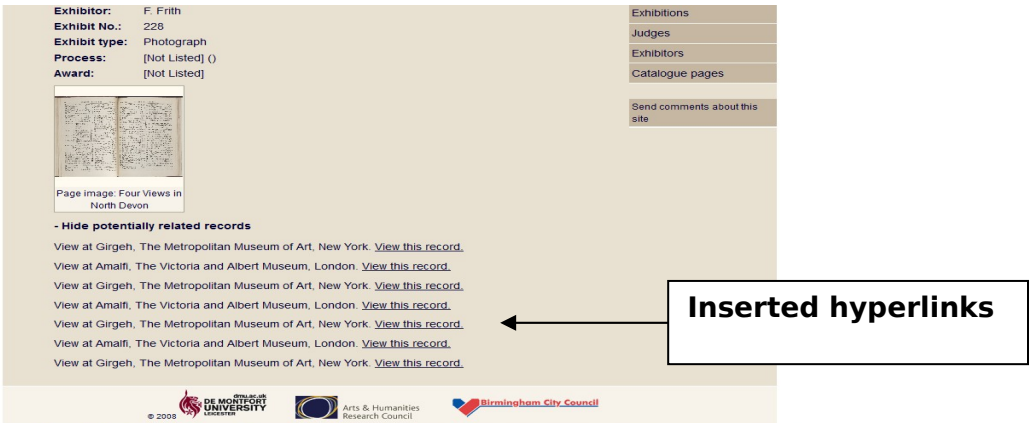
Figure 1. Parent child relationship between object records held by different partners.

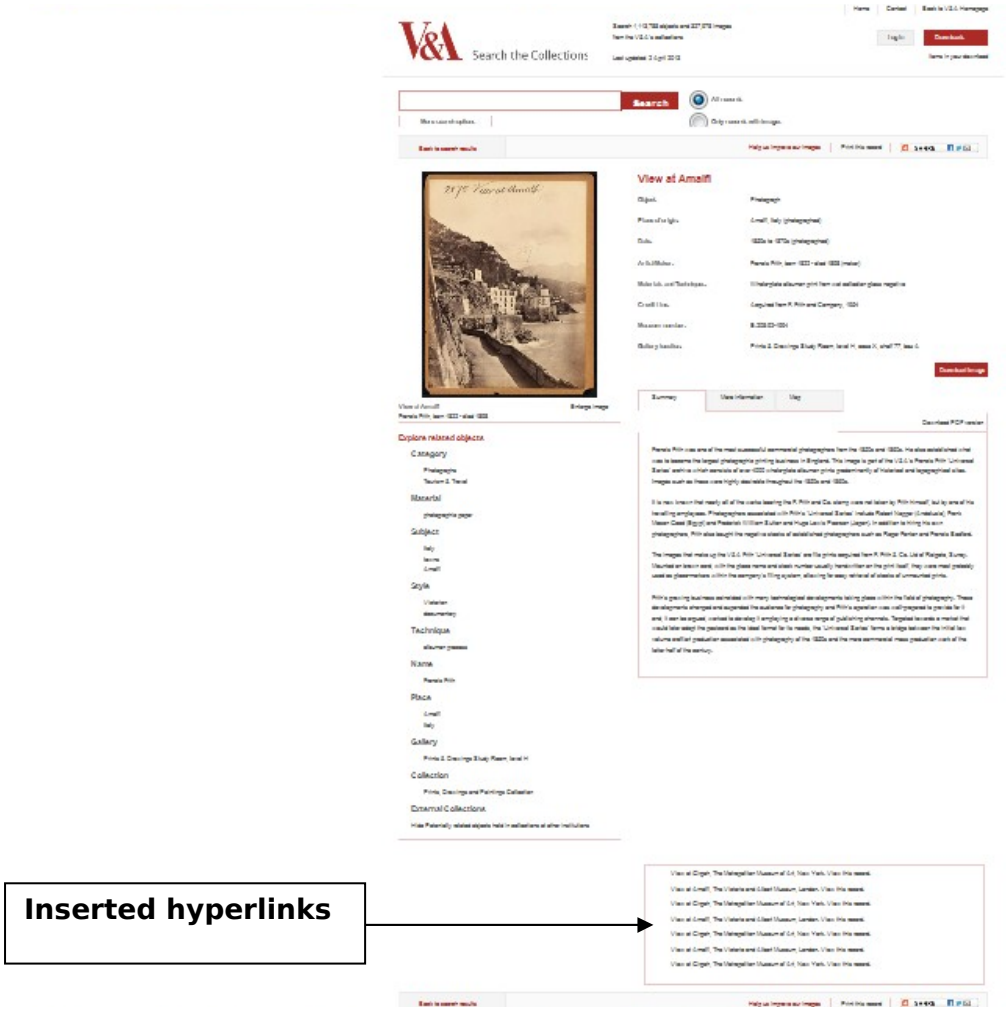
If there are no suggested matches, no hyperlinks will be shown.

What the widget looks like

The basic appearance will be a frame within the partner's webpage containing explanatory text and the suggested links. This frame can be bordered, using fonts and colours to distinguish it from the host page, or it can be borderless and treated graphically to blend in with the host page. The widget can be programmed to appear automatically if there are suggested links to a particular object, or it can be hidden initially, viewable only on selection of a toggle button that displays/hides the frame. Like the frame itself, this button can be programmed to appear only if there are links to display. Results will be displayed In blocks of up to 10, closest matches first. Users will be able to widen the parameters of their search by requesting more blocks of results.

The precise appearance of the widget can be tailored by, for example, linking it to the partners' web site CSS. See the examples below for the Exhibitions of the Royal Photographic Society 1870-1915 and the V&A web sites.





How it works

The widget uses the HTML i-frame (invisible frame) element to create the frame containing suggested hyperlinks. When a visitor opens a partner web page containing an object record, the (invisible) embedded code will query the FuzzyPhoto links database and retrieve any matches it finds there. If matches are found these will either be made immediately visible in the frame or a button will appear inviting visitors to request to view the links. (This is a configurable feature to allow partners to choose how the links appear). If the visitor selects one of the hyperlinks, a pop-up window will open in front of the partner web page. This window will provide a direct view of the suggested related object record as displayed on the relevant partner website.

Implications for partners

Branding

Is the frame required to blend in with, or stand out from the partner website? Partners may wish to integrate the i-frame visually to maintain visual harmony and corporate branding. Alternatively they may wish to emphasise that the i-frame content is third-party for which they are not responsible, by making its appearance radically different. This aspect is configurable using CSS.

Consistency versus redundancy

Consistency of appearance, navigation and behaviour helps visitors to understand and navigate websites and builds trust. This might be an argument for including the i-frame or its toggle button on every object record. However, matches are unlikely to be found for most records, so for most pages the frame/button would be redundant, requiring a message such as "no matches were found for this record". Inviting a visitor to follow a link, only to tell them that the link goes nowhere is bad web design. It is recommended therefore that the frame or its button is only displayed where there are suggested links, even though this will mean that web pages are not entirely consistent. This may be overcome by explanatory text in the frame.

Widget rubric

Effective web design is self-explanatory and should require no user instructions. However when a new user views a frame of suggested links some brief explanation may be necessary:

- To explain the third-party nature of the frame content, including a link to the Fuzzyphoto project website.
- To explain that links only appear throughout the site against object records where possible matches have been identified.
- To explain that following a link will open a new pop-up window.

This aspect of the widget will be tested via user trials to determine the most appropriate rubric.

Quality

Public perception of the partner website may be compromised if the widget performs erratically or produces poor matches. Erratic performance may be due to server loading and firewall issues. Technical tests will be conducted and results discussed with partners before the service is implemented publicly. However tests will have to be carried out on live partner websites and visitor perceptions will need to be managed appropriately during the testing period. Poor matches could result from setting the search parameters too wide (this is configurable so it should not be a problem).

Resilience

All data is backed up across multiple servers and to further persistent systems to ensure that in the case of data loss or server failure, service can be maintained. Virtualisation of the server has been employed for easy replication / backup of the web server. The data warehouse holding partner records is physically isolated from the internet and the links database/web server is protected behind the De Montfort university firewall.

Security

Security could be an issue if the widget could be used by malicious third parties to compromise partner web sites. The i-frame element has been identified as a potential security risk, but only for a site that is embedded inside an i-frame on a hostile site**.** Alternatively an i-frame element may be a security risk if any page on the site contains an XSS vulnerability which can be exploited, whereby an attacker can expand the XSS attack to any other page within the same domain that can be persuaded to load within an <iframe> on the page with XSS vulnerability. This is because content from the same origin (same domain) is allowed to access the parent content DOM (practically execute JavaScript in the "host" document). The advised defences against this type of attack are to add HTTP header X-Frame-Options: DENY and/or always correctly encode all user submitted data (that is, never have an XSS vulnerability on the site).

Sustainability

Capacity of the data warehouse and links database exceeds current usage allowing further expansion as required. A batch loading tool is being developed that will be made available at regular fixed intervals (monthly) to ingest new or revised catalogue record updates from partners. Partners wishing to submit revised catalogue records will be required to upload (FTP) them to the batch loader. Required data input format will be CSV in a format specified by DMU's FuzzyPhoto team. Newly ingested records will be added to the data warehouse and the search algorithms will output new suggested links to the links database. As soon as these updates have been made the new suggested links will be available to visitors at partner web sites. The FuzzyPhoto server cluster will be maintained by the De Montfort University Photographic History Research Centre through a service level agreement with De Montfort University Information Technology and Media Services (ITMS).